ESSAYS ON HEALTH CARE MARKETS

by

Matthew Noel White

A dissertation submitted to Johns Hopkins University
in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
May 2014

ABSTRACT

Chapter 1 examines how investments in health, through spending on preventive or curative care, affect subsequent spending on medical care among the retired population. I estimate a structural model of the "retired life cycle" using data from the Health and Retirement Study on single retired Americans. The traditional dynamic consumption-savings model is augmented by including two medical care goods, health investment and medical consumption, allowing the former to influence health and thus future need for the latter. After estimating, I conduct policy counterfactuals to ascertain the effect of a subsidy on preventive care on health outcomes, the total demand for medical care, and public expenditures on medical care for the elderly. A subsidy targeted at healthy, lower income individuals improves longevity but does not reduce total demand for medical care over the remaining lifetime of the targeted population.

Chapter 2 uses the estimated model to investigate how medical inflation affects the consumption, medical care, and saving decisions of retired Americans. I simulate the behavior of young retirees under the average rate of medical inflation and two alternative rates. Further, the effects of medical inflation are decomposed into two parts: the intratemporal or reactionary effect, motivated by experiencing changes in the current price of care; and the intertemporal or precautionary effect, arising from foreseeing changes in future price growth. The decomposition shows that changes to consumption represent a balanced tension between opposing precautionary and reactionary effects, while the effect on medical care is driven almost entirely by the reactionary effect.

Chapter 3 presents a model to analyze consumer welfare, price, and competition in a three-way market among patients (consumers), medical providers, and insurers. Examples are used to demonstrate that consumer welfare is not necessarily increasing in the number of insurers, but instead exhibits a U-shaped pattern. While insurers compete with each other for customers, they also act as collective bargaining agents on behalf of patients in determining the equilibrium price of health care with providers. The entry of an additional insurer thus has contradictory effects on prices and consumer welfare, reducing prices through competition but increasing them through reduced bargaining power of incumbent insurers. Moreover, the more favorable contracts allow individuals to purchase care more often, shifting out the demand curve for care and resulting in a higher equilibrium price. I find the equilibrium of the game under all combinations of monopoly and competition between providers and insurers, and present examples to analyze the effects of insurer entry on consumer welfare and the price of care.

READERS/ADVISERS

This dissertation was supervised by Hülya Eraslan and Christopher Carroll. Helpful comments were provided by Robert Moffitt, Nicholas Papageorge, and Elena Krasnokutskaya. I would like to thank my oral defense committee members: Hülya Eraslan, Christopher Carroll, Nicholas Papageorge, Bradley Herring, and Angelo Mele (with alternates Yuya Sasaki and James Fill, just in case).

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

Chapter 2

Chapter 3

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 3

This page intentionally left blank.

Chapter 1:
An Ounce of Prevention at Half Price:
Evaluating a Subsidy on Health Investments

Abstract

This chapter examines how investments in health, through spending on preventive or curative care, affect subsequent spending on medical care among the retired population. I estimate a structural model of the "retired life cycle" using data on single retired Americans from the Health and Retirement Study, including a measure of health constructed from objective and subjective data. The traditional dynamic consumption-savings model is augmented by including two medical care goods, health investment and medical consumption, allowing the former to influence health and thus future need for the latter. After estimating, I conduct policy counterfactuals to ascertain the effect of a subsidy on preventive care on health outcomes, the total demand for medical care, and public expenditures on medical care for the elderly. A subsidy targeted at healthy, lower income individuals improves longevity by 0.76 months at a public cost of $760 per capita. However, the subsidy does not reduce total demand for medical care over the remaining lifetime of the targeted population. The government ultimately pays for the policy twice: first for the subsidy directly, and again when individuals use the savings to finance more medical consumption, partially paid by Medicare.

Notes: Many thanks are owed to Hülya Eraslan and Chris Carroll for their constant feedback and support. Helpful comments were also provided by Robert Moffitt, Elena Krasnokutskaya, and Nicholas Papageorge. This study makes use of data from the Health and Retirement Study (HRS); the HRS is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan.

1 Introduction

It is well established that wealthier individuals live longer and have better health (Deaton (2002)); and conditional on current health, richer individuals experience slower health deterioration (Case and Deaton (2003)) and spend more on medical care. It is likewise uncontroversial that, all else equal, healthier people have lower medical expenses (Thorpe, Florence, and Joski (2004)). These facts suggest an intuitive story: Those with more resources are better able to invest in their health through medical care, leading to lower future health expenditures than they would have in the absence of these health investments. This in turn raises the possibility that lowering the cost of health investments could motivate individuals to better preserve their health. They would thus need less medical care to manage disease in the future, potentially resulting in a net decrease in total demand for medical services. The Patient Protection and Affordable Care Act includes such subsidies, significantly increasing the generosity of Medicare's coverage for preventive care. President Obama has asserted that this provision "provid[es seniors] the kind of preventive care that will ultimately save money throughout the system."[1]

An economic model suitable for evaluating this claim must capture a key intertemporal tradeoff between medical care in the present and future: Additional purchases of preventive care improve the distribution of health in the population, which in turn reduces the need for care to mitigate or cure future ailments. Existing models treat medical spending either as an investment in the stock of health (as in the classic Grossman (1972) model) or as a stochastic shock to wealth (e.g. DeNardi, French, and Jones' (2011)), but both models must be combined to address the dynamic effects of a subsidy on health investments.

To this end, this paper uses an estimated structural model of the "retirement life cycle" that incorporates these features to gauge the effects of a subsidy on preventive care. The model includes two medical care goods: "health investment", which affects the subsequent stock of health; and "medical consumption", the need for which depends stochastically on health. This two-good structure creates the intertemporal tradeoff in medical care purchases. I estimate the model with the simulated method of moments using data from the Health and Retirement Study (HRS) to match conditional profiles of assets, medical expenses, health, and mortality. To evaluate the effects of a subsidy on preventive care, I simulate the estimated model with a government subsidy on the investment-type medical care good and compare the outcomes to a baseline simulation.

[1] First Presidential debate, October 3, 2012. Obama has made similar claims on other occasions.

While the subsidy's direct benefit is improved health and longevity, it also has the potential to be cost-saving for the government. The possibility of a net negative cost— savings from lower future reimbursements exceeding the direct cost of the subsidy— is maximized when it is restricted to lower income households. The welfare of the poorest individuals is maintained through government transfers (e.g. Supplemental Security Income, Medicaid) when their own resources are insufficient to attain a minimum standard of living. Because they do not expect to be able to achieve utility much higher than the statutory minimum, these individuals effectively face very little risk in their utility. This low utility risk diminishes the incentive to preserve or improve the health stock to guard against medical expenses, as very little of these costs will be borne out-of-pocket. Poor individuals' current low level of health investment means that there are significant health gains to be made with additional investment. Moreover, the financial benefits of reducing their future medical care costs would be realized largely by the government, which bears most of their medical expenses through social safety net programs. In this way, the subsidy is most efficient for the government in both cost and benefit when applied to poorer individuals.

I find that the hypothetical subsidy is not successful in reducing government expenditures, but it does improve longevity by an average of 0.76 months for the targeted population: individuals in the bottom two income quintiles in sufficiently good health. The subsidy adds about $760 in government spending per targeted individual over their lifetime, or $167 per individual when spread out over all retirees. It minimally reduces out-of-pocket medical spending, but increases total lifetime medical spending by $979 per targeted individual, or $217 across all retirees. The increase in government expenditures and total medical spending occurs because the subsidy frees up resources for individuals to spend on consumption or medical consumption. In this way, the subsidy largely causes a shift in the composition of out-of-pocket medical spending without changing its magnitude significantly. Meanwhile, the government pays directly for the subsidy and indirectly through larger Medicare reimbursements for medical consumption.

1.1 Discussion

My model employs two medical care goods– "medical consumption" and "health investment". Medical consumption represents *mitigative care*, which is used to relieve or manage pain, disability, or symptoms associated with an illness (e.g. morphine). It contributes to the patient's current utility but does not improve future health. Health investment represents both *curative care* and *preventive*

care. The former is employed to treat the illness itself and thus avert future pain, disability, or symptoms (e.g. penicillin). The latter is used to prevent an illness from arising or progressing in the first place (e.g. the flu vaccine). Thus, they both act like an investment good by influencing the underlying health stock: preventing degradation in good health or increasing the stock in bad health. In my model, an individual's "medical needs" are randomly drawn each period, determining how much medical consumption is valued relative to consumption of other goods and services. Because the distribution of medical needs depends on the health state, current health investment reduces future medical consumption– the critical intertemporal tradeoff.

The key parameters of the model that determine how simulated individuals react to the subsidy are identified by matching median health profiles as they vary by income and assets. Conditional on current health, wealthier individuals have better future health outcomes. My model explains these differential patterns through endogenous variation in the level of health investment purchased: wealthier individuals have higher consumption levels and thus lower marginal utility of consumption, inducing them to seek utility by investing more in their health to attain more periods of life. As in Arcidiacono, Sieg, and Sloan (2007), low or no health investment by poorer individuals can be optimal if their expected utility in subsequent periods is not sufficient to justify sacrificing consumption for the sake of extending life. The efficacy of the health investment good in producing health (in both level and curvature) and the absolute utility of life are identified by variation in the rate of health decline across income and asset quintiles. A more complete discussion of identification can be found in section 4.2.

By explaining variation in the rate of health decline across income and asset levels through different levels of health investment, my model takes a strong stand on a phenomenon that could have other origins. For example, wealthier individuals could have a lower discount rate, which motivated them earlier in life to invest in human capital to generate their high income and now causes them to forgo unhealthy behaviors, extending their lives. Moreover, the causality could be reversed: unhealthy individuals frequently missed work, preventing productivity growth and keeping their incomes low. Related, the nature of low-paying jobs might have affected workers' bodies so that health is not only lower at retirement but also deteriorates faster.

These possibilities are inconsistent with other patterns in the data or are insufficient to account for the magnitude of the differential in health outcomes. If the first alternative explanation were true, then the effects of a lower discount factor should also be seen in the rate of asset depletion for low income individuals, as they are less inclined to reserve resources for the future. To the contrary, the

estimated model matches poor individuals asset profiles better than it does for wealthier individuals. The latter two explanations cannot be addressed without deeper data on the work histories of HRS respondents. While illness-related interruptions in work histories certainly affect income, they are unlikely to be a major contributor the large income inequality observed among both the working and retired population.

The model used in this paper does not capture all features relevant to the determination of total demand for medical care. A more complete model would account for multi-person households, as in husband-and-wife bargaining models (e.g. Blau and Gilleskie (2006)), to address the sharing of resources and joint optimization. Moreover, the model ignores the role of other health-related behaviors– such as drinking, smoking, and exercise– considered by others in the literature (e.g. Khwaja (2010)). In addition, the assumptions of my model make it suitable only for analyzing the retired population, for whom income risk is low and does not depend on current health. Nonetheless, the model is a building block toward these more inclusive models of health and medical expenses, as my model of a single retired individual represents the end-state both for working singles and for retired couples. Further, data on these behaviors are provided in the HRS, so the model and estimation can be extended to account for these effects. Despite these shortcomings, my model represents one of the first to employ a two-good approach to medical expenses, separately addressing the consumption and investment aspects, and thus combine the two major strains of the literature.

1.2 Literature

My model draws from two lines of literature on medical expenses, each of which are insufficient on their own to capture the intertemporal tradeoff between health investment and medical consumption. Literature that treats medical spending as an investment in a stock of "health capital" allows health to be endogenous to individuals' decisions, but can only match the large observed variance of medical expenses with an unrealistically large variance of health shocks. On the other hand, literature that considers the role of medical expenses in motivating saving by the elderly can match the extreme variance of medical spending, but treats health as exogenous and thus cannot explain differences in health outcomes across income levels. When these models are combined, health investment in the present can be used to shift down the distribution of other medical expenses in the future.

Economic models of health investments can be traced back to Grossman's 1972 model in which agents' optimal stock of health declines as they age due to the rising costs of maintaining health.

A large empirical literature has spawned from the original model, but the functional form of health investment in my model is attributable to Yogo (2009).[2] Yogo models health as a continuous variable that evolves according to normally distributed shocks, conditional on current health and log medical spending. While the model is used to explain differences in mean medical spending by health and other covariates, his model does not address the large variance of medical spending within any given group. In his model, any individual variation in medical expenses is driven by shocks to the health state (whose variance is calibrated from the data), so the simulated variance of expenses is likely much too small.

A recent working paper by Ozkan (2010) uses a model with two medical care goods to explain the profiles of medical spending by wealthy and poor individuals, specifying both medical care goods as investments in two stocks of health capital. Physical capital determines survival probabilities, while preventive capital governs the distribution of shocks to physical capital; neither type of care can improve the stock, merely offset current period losses. The health capital stocks are unobservable abstractions, thus the estimation does not attempt to match rates of health decline, only longevity and medical expenses. Moreover, because the large variance of medical expenses depends entirely on the magnitude of shocks to physical capital, Ozkan's model unrealistically predicts that after controlling for income levels, people of lower assets are much more likely to die suddenly, as they are unable to guard against sudden large health shocks. In contrast, my estimation treats health as an observable quantity and matches both health deterioration rates across wealth and income and distributions of medical spending by health. Moreover, the high variance of medical expenses does not imply an unrealistic distribution of health shocks in my model, as stochastic need for medical consumption drives much of the medical expense variance.

The structural model is most directly based on DeNardi, French, and Jones (2010) (hereafter "DFJ"), which explores how large and volatile medical expenses motivate saving among the elderly. The model particularly draws from their specification in which the quantity of medical care is determined endogenously rather than as an exogenous shock. DFJ are able to match asset and medical spending profiles of the very old, and they find that the serial correlation of medical needs shocks is a major motivation for the savings maintained by even wealthy retired individuals. My model mirrors the structure of DFJ, but adds a second medical care good that affects the health state transition rather than treating income as a direct input into health evolution. Further, I capture the serial correlation of medical needs shocks through persistence of a continuous health variable, in

[2] Yogo's paper has since been updated to use a different functional form.

contrast to DFJ's specification of binary health with underlying permanent shocks to medical needs.

Related models also do not focus on the tradeoff between present medical spending to invest in health and future medical spending to mitigate illness, as they treat medical care as a single good or omit savings. Both French and Jones (2005) and Blau and Gilleskie (2008) investigate how retiree health insurance benefits— the ability of a worker to continue to purchase medical insurance from an employer after retirement— affects the decision to retire before reaching Medicare age. Blau and Gilleskie model the medical care decisions of men approaching retirement as they make discrete decisions about work and medical care utilization. The single medical care good affects both utility and the health state transition probabilities. Khwaja (2010) estimates a discrete choice model of consumption, medical expenses, and health-affecting behaviors (smoking, alcohol consumption, exercise) to estimate the willingness to pay for Medicare. Assets and saving behavior are not modeled, with individuals assumed to consume any remaining income after paying insurance premiums and out-of-pocket medical costs. While medical care again provides direct utility and affects health evolution, it is modeled as a unitary good. DePreux (2011) considers how Medicare can generate an anticipatory or ex ante moral hazard effect on the same health behaviors considered by Khwaja, estimating a reduced form based on a discrete choice model that also does not include a savings decision.

In summary, existing models of medical spending treat the health state as exogenous or do not allow need to vary stochastically. The models that allow medical care to affect utility and the production of health tend to combine these effects into a single good, treat medical care as a discrete choice, and do not model asset accumulation and decumulation. Moreover, many models specify health as binary ("good" or "bad") and thus cannot account for marginal improvements in health. My model combines these threads to incorporate both the consumption and investment aspects of medical services in the presence of asset accumulation and decumulation, simultaneously matching the distributions medical spending and health evolution. While it omits the role of non-medical behaviors and the inclusion of the working population, it is nonetheless a step forward from existing models.

The medical literature is somewhat pessimistic about the cost efficiency of preventive services. Near universal application of a preventive service (a screening test, say) will often have costs that outweigh its benefits because the majority of recipients will test negative, offering no medical benefit; moreover, because type II errors are much more costly, these tests will tend to generate many false positives, resulting in further costs. The scientific consensus is that while some universal applications

of preventive care (such as childhood vaccinations) reap significant savings, the majority of services studied so far do not (Russell (1993, 2007, 2009)). Intuitively, preventive interventions are more cost efficient when more narrowly targeted at the most at risk population, extracting the most expected benefit from each application. If a wide array of preventive services are subsidized for a population that has low utilization rates, the individuals who receive a service as a result of the subsidy are those who were marginal between its costs and benefits before the subsidy and thus the most efficient group for each treatment.

The remainder of the paper proceeds as follows: Section 2 specifies a dynamic model of consumption and savings with two medical care goods; Section 3 describes the HRS data and transformations thereof used in the estimation; Section 4 describes the SMM estimation method and identification strategy and presents parameter estimates; Section 5 conducts counterfactuals of the subsidy on preventive care; and Section 6 concludes.

2 Model

In this section, I build a consumption-savings model with two medical goods. It is intended to capture the relevant behavior of single, retired individuals over the age of 65 living in the United States. Many of the model features are based on the "endogenous specification" of DeNardi, French, and Jones (2011) (DFJ), but it also includes a health investment good to endogenize health. This section will specify each of the major components (including the utility function, the distribution of medical needs, and the distribution of the subsequent health state) in turn before moving on to a description of its solutions and a discussion of the modeling choices.

2.1 Specification

Individuals in the model, indexed by i, represent unmarried retired persons over the age of 65 living in the United States. Individuals are lifetime expected utility maximizers over a finite lifetime, with a common utility function and an intertemporal discount factor δ over discrete time t (in two-year periods). At time t, individual i has a real income y_{it} that is received conditional on survival to that period. Individual i has a stock of health $h_{it} \in [0, 1]$ at time t, where 0 represents being dead and 1 represents "perfect" health. Individual i's net real assets are denoted by $b_{it} \geq 0$.

Individuals purchase non-negative quantities of three goods each period: composite consumption c_{it}, medical consumption μ_{it}, and health investment κ_{it}. All individuals face the same relative price

p_t of medical care at time t; the price of the composite consumption good c is normalized to one. I assume the sequence $\{p_t\}_{t=0}^{\infty}$ is exogenously given, and that all individuals have common and correct beliefs about the future path of p_t. Note that t should not be thought of as age, but as time itself.

Sequence of events: Each period, the sequence of events for a living individual is as follows. The individual receives income and pays (exogenous) medical insurance premiums, then his medical needs shock is realized. Given the shock, he chooses optimal quantities of goods subject to his budget constraint. Finally, his health state evolves and the next period begins.

Utility function: The utility flow of living individual i at time t depends on composite consumption c_{it}, medical consumption μ_{it}, as well as his health state h_{it} and the realization of his medical needs shock η_{it}. An individual who has died since the previous period receives a terminal payoff based on his assets, acting as a bequest motive. I assume all individuals have a common utility function given by:

$$u(c_{it},\mu_{it};\eta_{it},h_{it},h_{it-1},b_{it}) = \begin{cases} (1+\alpha_1 h_{it})\frac{c_{it}^{1-\rho}}{1-\rho} + \eta_{it}\frac{\mu_{it}^{1-\nu}}{1-\nu} + \varsigma_0, & \text{if } h_{it} \in (0,1] \\ \omega_1 \frac{(b_{it}+\omega_0)^{1-\rho}}{1-\rho}, & \text{if } h_{it}=0 \text{ and } h_{it-1}>0 \\ 0, & \text{if } h_{it}=0 \text{ and } h_{it-1}=0. \end{cases} \quad (1)$$

Medical consumption is simply a second consumption good with a random marginal utility determined by the realization of the medical needs shock η_{it} (see below). The coefficient of relative risk aversion for medical consumption ν is restricted to be greater than one so that larger medical needs are a penalty to utility; this adverse effect is mitigated through medical consumption. The utility level shifter ς_0 represents the value of life and allows utility to be greater than zero even if risk aversion for composite consumption $\rho > 1$; without this term, individuals would never seek to extend life through health investment, as death (utility zero) would be preferable. The marginal utility shifter for health α_1 is included to allow non-separability between health and composite consumption (Finkelstein et al (2008)), which may be useful in explaining the patterns observed in the data across income and asset levels. The bequest motive in the second line of (1) has flexibility in both curvature (ω_0) and scale (ω_1) and is identical to that used by DFJ.[3]

Medical needs distribution: The medical needs shock is drawn from a Weibull distribution whose

[3] Arguments of the utility function after the semicolon will be suppressed in the equations that follow.

scale parameter depends on the individual's sex, age, and health and shape parameter is a linear function of health. Formally:

$$\eta_{it} \sim f(\eta|Z_{it}) = \frac{k_{it}}{\lambda_{it}} \left(\frac{\eta}{\lambda_{it}}\right)^{k_{it}-1} e^{-(\eta/\lambda_{it})^{k_{it}}}, \qquad k_{it} = \beta_{k0} + \beta_{k1} h_{it}, \tag{2}$$

$$\lambda_{it} = \exp\left(\beta_0 + \beta_s sex_i + \beta_{a1} age_{it} + \beta_{a2} age_{it}^2 + \beta_{h1}(1-h_{it}) + \beta_{h2}(1-h_{it})^2\right), \tag{3}$$

where $Z_{it} = (sex_i, age_{it}, h_{it})$ is a vector of characteristics of individual i which are exogenous at time t. I will collectively refer to the parameters β as the shock parameters. This form allows for medical needs to become both greater on average and have a greater variance as health declines without restricting the relationship between the mean and variance, and fits the data better than the lognormal form often used (see Section 2.2).

Health state distribution: At the end of each period, a living individual first experiences a mortality shock, modeled as a probit based on sex, age, and health. Individuals who survive the mortality shock then receive a draw of their subsequent health state, the distribution of which depends on their age, sex, health, and health investment. This process is described formally by:

$$h_{it+1} = \begin{cases} \max\{0, \min\{1, \hat{h}_{it+1}\}\}, & \text{if } \Theta_{it} \leq 0 \text{ and } h_{it} > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

$$\Theta_{it} = \theta_0 + \theta_s sex_i + \theta_{a1} age_{it} + \theta_{a2} age_{it}^2 + \theta_{h1}(1-h_{it}) + \theta_{h2}(1-h_{it})^2 + \epsilon_{it}^{\theta}. \tag{5}$$

$$\hat{h}_{it+1} = \gamma_0 + \gamma_s sex_i + \gamma_{a1} age_{it} + \gamma_{a2} {age_{it}}^2 + \gamma_{h1} h_{it} + \gamma_{h2} h_{it}^2 \tag{6}$$

$$+ \exp(\gamma_{m1} + \gamma_{m2}(1-h_{it}))(\log(\kappa_{it} + \exp(\gamma_{m0})) - \gamma_{m0}) + (\sigma_0 + \sigma_1(1-h_{it}))\epsilon_{it}^{h}.$$

When $h_{it} = 0$ and the individual is already dead, the next health state is also death. Both the mortality shock ϵ_{it}^{θ} and the health shock ϵ_{it}^{h} are distributed standard normal. Note that individuals who survive the mortality shock might still die if they receive a shock to health that would result in $h_{it} \leq 0$; next period's health is also censored above by "perfect health" of one. The distribution of future health generated by (4), (5), and (6) will be denoted by $g(h_{it+1}|\kappa_{it}, Z_{it})$.

The first line of (6) represents the mean of subsequent health conditional on age, sex, and current health. The first term in the second line of (6) represents the contribution of health investment κ_{it}. The first term of the second line of (6) is the contribution to future health from health investment–

the "health production function". The first factor, $\exp(\gamma_{m1} + \gamma_{m2}(1 - h_{it}))$, is the efficacy of health investment at the current health state. γ_{m1} is the base efficacy, while γ_{m2} represents the differential efficacy of curative care relative to preventive care, as an individual moves into poorer health. The form of the second factor, $(\log(\kappa_{it} + \exp(\gamma_{m0})) - \gamma_{m0})$, allows the curvature of health production to vary with γ_{m0}: a lower value of γ_{m0} means that the marginal returns to health investment decrease more rapidly. Note that when $\kappa_{it} = 0$ (there is no health investment), the health production function returns zero. The second term of the second line of (6) allows the health shocks to have higher variance when the individual is sicker.

In order to make feasible the solution of optimal behavior over the retirement life cycle, I assume that there is some maximum age beyond which individuals cannot live. If an agent survives to this age, then $g(\cdot)$ becomes degenerate with a unit mass at $h_{it+1} = 0$.

Budget constraint: The individual must choose an affordable bundle subject to his budget constraint, which requires next period's assets to be non-negative. Welfare payments w_{it} are provided to individuals who cannot achieve at least some utility floor $\underline{u}$ (set by government policy) with his available resources. When $w_{it} > 0$, it is equal to the dollar value that will exactly allow the agent to reach the utility floor if he does not purchase any health investment, with his assets for the next period set at zero. The individual pays exogenous insurance premiums z_{it} and only pays a fraction q_{it} of total medical costs, his copay rate. His budget constraint is thus:

$$b_{it+1} = R_t \left(b_{it} + y_{it} + w_{it} - z_{it} - c_{it} - q_{it} p_t(\mu_{it} + \kappa_{it})\right) \geq 0. \tag{7}$$

Individual's problem: The individual's problem in each period is to maximize his expected lifetime utility by choosing the optimal quantities of c, μ, and κ while adhering to his budget constraint. That is, he must balance the immediate utility of composite and medical consumption against the future payoff from health investments and the need to save assets as a buffer against future uncertainty in medical needs and health.

Assume individual i's value function for period $t + 1$, $V_{it+1}(b, h)$, is known, representing the expected lifetime value of beginning that period with given levels of assets and health. Once i

realizes his medical needs shock of η_{it}, he faces the following problem:

$$V_{it}^{\bullet}(b_{it}, h_{it}, \eta_{it}) = \max_{c_{it},\mu_{it},\kappa_{it}} \left[u(c_{it}, \mu_{it}; \cdot) + \delta \int_0^1 V_{it+1}(b_{it+1}, h) g(h|\kappa_{it}, Z_{it}) dh \right] \text{ s.t. (7).} \quad (8)$$

The bullet denotes that this is the intermediate value function, or the expected lifetime payoff for i from receiving a medical needs shock of η_{it} after he enters period t with the particular level of assets and health. This can be integrated with respect to the shock distribution in order to find the value function at the start of the period, before the shock is realized:

$$V_{it}(b_{it}, h_{it}) = \int_0^{\infty} V_{it}^{\bullet}(b_{it}, h_{it}, \eta) f(\eta|Z_{it}) d\eta. \quad (9)$$

The previous two equations can be used to solve the life cycle problem backwards until initial Medicare eligibility at age 65, successively finding the optimal bundle of (c, μ, κ) at each point in the state space of (b, h, η) (that is, $\mathbb{R}_+ \times [0, 1] \times \mathbb{R}_+$), integrating across η to find the value function, and rolling back to the previous period.

Optimality conditions: Assuming an interior solution, there are three first order conditions for the maximization problem in (8), respectively for composite consumption, medical consumption, and health investment:

$$(1 + \alpha_1 h_{it}) c_{it}^{-\rho} - \delta R_t \int_0^1 V_{it+1}^{'}(b_{it+1}, h) g(h|\kappa_{it}, Z_{it}) dh = 0. \quad (10)$$

$$\eta_{it} \mu_{it}^{-\nu} - q_{it} p_t \delta R_t \int_0^1 V_{it+1}^{'}(b_{it+1}, h) g(h|\kappa_{it}, Z_{it}) dh = 0. \quad (11)$$

$$\delta \int_0^1 V_{it+1}(b_{it+1}, h) g_{\kappa}(h|\kappa_{it}, Z_i t) dh - p_t q_{it} \delta R_t \int_0^1 V_{it+1}^{'}(b_{it+1}, h) g(h|\kappa_{it}, Z_{it}) dh = 0. \quad (12)$$

Recognizing the commonality across the second term in these three equations, they can be combined into an intuitive quadruple equality, which says that the marginal benefit of a dollar spent on composite consumption, medical consumption, health investment, and savings must all be equal at the optimal bundle:

$$\begin{aligned}(1 + \alpha_1 h_{it}) c^{-\rho} = (q_{it} p_{it})^{-1} \eta_{it} \mu^{-\nu} = \delta (q_{it} p_{it})^{-1} \int_0^1 V_{it+1}(b_{it+1}, h) g_{\kappa}(h|X_{it}, h_{it}, \kappa_{it}) dh \\ = \delta R_t \int_0^1 V_{it+1}^{'}(b_{it+1}, h) g(h|X_{it}, h_{it}, \kappa_{it}) dh \equiv (\delta R_t) \, \mathrm{E}_t \left[V_{it+1}^{'}(b_{it+1}, h_{it+1}) \right]. \end{aligned} \quad (13)$$

Note that the leading factor in each expression is the inverse price of a unit of the good. The first and second expressions can be rearranged to provide a closed form solution for medical consumption based on the optimal quantity of composite consumption:

$$\boldsymbol{\mu}^*(c) = ((1 + \alpha_1 h_{it})pq)^{-1}\eta)^{1/\nu} \cdot c^{\rho/\nu}. \tag{14}$$

Thus the three-dimensional optimization problem is effectively a two-dimensional problem instead, as for any candidate quantity for optimal consumption, there is exactly one quantity of mitigative care that can be an optimum along with it. This greatly reduces the complexity of the problem and speeds up the computation of the solution.

2.2 Discussion

With the description of the model in hand, some discussion of the assumptions is warranted. Many of the choices were made to hew as close as possible to the model of DFJ to allow comparability of the results. For example, the utility function used in (1) is identical to theirs, with the addition of the level shifter ς_0. Because the value of death is normalized to zero and health is endogenous, the level shifter is necessary when ρ and ν are greater than one, as is realistic and has been estimated previously. In the absence of the level shifter for the value of life, the entire utility function would be negative and agents would make no efforts to extend their life, preferring the relatively higher payout from death. The utility floor and welfare transfers mimic those in the DFJ model and are meant to represent a minimum standard of living through social programs such as Medicaid and Supplemental Security Income (SSI). DFJ find that the utility floor is a key determinant of behavior, even when individuals have non-negligible assets. It provides insurance against catastrophic outcomes and thus reduces the precautionary saving motive.

The assumption of an exogenous, known, and certain path of income for each type of agent is justified by the retired status of the individuals considered in this model. Retired individuals derive most of their income from fixed sources such as pensions, annuities, and Social Security. The payments yielded by these sources were determined by the labor and investment decisions made by the individual during the working life and are considered exogenous in this model. Critically, the income of a retired individual does not depend on health, simplifying the analysis. In practice, each individual in the model is not given a unique timepath of income, but rather respondents in the data are grouped by sex, cohort, and income quintile into one hundred fifty "types", each with a

timepath of real income.

An innovation of this model over other recent work is the inclusion of a non-static relative price of medical care, p_t. Models that treat medical expenses as exogenous capture this effect through cohort effects or a time-varying distribution of cost shocks, but little work has been done that models the price explicitly or varies expectations about the price. Medical inflation (growth of the medical component of the CPI beyond the non-medical component) has been positive in all but one year since 1981, averaging 2.37% over the past thirty years. The trend is widely known and is expected to continue, and so its inclusion in a dynamic model of medical decision-making is significant.

The Weibull form used in (2) is motivated by visual inspection of the distribution of total medical care costs across the sample as well as subsamples conditional on health status. The CDF of these distributions appears to be Weibull, with a shape parameter which varies with health and is less than one. This is a departure from most of the literature, which uses a log-normal form for medical expenses. The choice of a log-normal distribution seems to be motivated less by the fit to the data and more by the ease of specifying a traditional serial correlation structure. In my model, the serial correlation is instead subsumed into the continuous health state, which is much more coarse in other models (often binary). In this way, what others label as a persistent needs shock or unobserved heterogeneity is observable in this model. It is worth noting that the specification of k varying with health is mostly interchangeable with one in which k is constant but ν varies with health.

The structure of the health distribution $g(h_{it+1}|\kappa_{it}, Z_{it})$ is likewise necessitated by the data. While poor health is certainly related to higher mortality, ostensibly healthy individuals commonly die suddenly due to acute conditions such as a heart attack or stroke.[4] Without the separate mortality shock, a very large variance of health shocks σ_0 would be necessary to match observed levels of mortality among healthy older individuals. This would not be congruent with observed health variance within individuals across time. The two-shock structure thus allows the estimated model to match both mortality rates and the volatility of health. Moreover, a dual probit analogue is commonly used in models that use discrete health states.

3 Data

Estimating the parameters of the model described in the preceding section requires data on individuals' income, assets, health, and medical expenses. The Health and Retirement Study (HRS),

[4]With HRS data only collected every two years, these deaths need not be particularly sudden to appear so in my data.

a panel study of several thousand older Americans as they transition into retirement and old age, provides these data in two year waves from 1992 through 2010. The HRS is primarily concerned with the dynamics of labor supply, income, assets, and various aspects of health and medical care utilization. Data for this project are derived from eight waves of the HRS, from 1996 to 2010.

3.1 HRS Data

The sample for my estimation is restricted to retired individuals over the age of 65 who are unmarried and whose total net assets in the first (otherwise qualifying) data wave in which they appear are strictly between -$3000 and $8,000,000 (in year 2000 dollars), with negative assets censored to zero. The restriction to the retired population is necessitated by the assumption that each individual has a known and fixed path of income, conditional on survival. Likewise, married respondents are not included because their maximization problem is complicated by the presence of other individuals in the household, leading to questions about resource sharing and joint utility maximization. The assets restrictions are simply a matter of convenience to bound the values of wealth at which the behavioral functions are calculated, and represent approximately the middle 98% of the wealth distribution.

The RAND Corporation has combined the rich data on assets, liabilities, and income to produce estimates of each individual's total income and total net wealth, imputing missing values where necessary. These assigned values are used in my estimation. All asset and income values are deflated using the CPI to transform them into year 2000 dollars. As seen in Table 3, both income and assets have very long-tailed distributions: about 15% of individuals have no assets, while about 4% have over $1,000,000. For numerical convenience, the model operates in units of $10,000. As with assets and income, the total dollar value of medical care is deflated by the CPI as well as the relative price of medical care for that period (the medical component of the CPI divided by the non-medical component) to construct the quantity of medical care in time-invariant units. For future periods, I assume that the relative price of medical care grows at the historic average rate (see Figure 1).

Because solving the model for every individual in the sample is both computationally infeasible and impossible given the data (unobserved income beyond the period of actual death), I instead categorize each respondent into one of one hundred fifty "types" based on their sex, cohort (birth years 1910 to 1939 in two year blocks), and income quintile. The income quintile is determined by the relative rank of the discounted sum of the first two waves of income data for an individual relative to his sex-cohort peers. To construct income profiles for each type, I calculate median income for

each type in each period, filling low-observation values using adjacent periods or similar types (see appendix). Because the data are very sparse at higher ages, the paths of income beyond age 95 are assigned somewhat arbitrarily. Neither real nor simulated agents often survive to these ages, so the assumed income levels are likely harmless to the estimation.

3.2 Constructed Variables

Continuous health measure: I construct a novel measure of health continuous on the unit interval using individuals' self reports of subjective and objective health conditions. The HRS asks respondents about physical condition (severity of regular pain, days spent in bed, days with lost urine, etc), past and current medical conditions (cancer, stroke, etc), mobility (ability to walk, bend over, pick up a dime, push a chair, etc), and capacity for conducting daily activities (dressing, using the phone, taking prescriptions, managing money, etc), as well as a categorical subjective evaluation of their overall health (five categories from poor to excellent). To construct a continuous measure of health, I estimate an ordered probit of the subjective health category on a large number of objective health measures, somewhat similar to Bound, Stinebrickner, and Waidman (2010). This generates weights on each of the objective measures that determine their relative importance in subjectively labeling health. The fitted values from the ordered probit (see Table 2) are linearly transformed into the unit interval, with the top cutoff point translating into a health value of 1 while the third percentile of the fitted values below the lowest cutoff translating into a health value of 0.01.

The results of the health status ordered probit are reported in Table 2. Nearly all of the objective conditions are significant at the 1% level or better, and almost as many have the expected sign. Additional estimations (not reported here) reveal that the coefficients for the medical conditions vary somewhat between men and women when estimated separately, with men's subjective health usually slightly more sensitive to a particular condition, but not so different as to warrant separate equations. Similarly, subjective labeling of health status does not seem to depend on age; when the critical points of the ordered probit are re-estimated on subsamples by age (holding the coefficients fixed), they do not significantly vary between the ages of 65 and 90. When sorted by reported subjective health status, the distributions of the constructed health value reveal that the estimation does an adequate job of separating the categories.

Insurance functions: The budget constraint (7) includes insurance premiums and a copay rate for the individual. Following DFJ, I estimate simple insurance functions to assign these values to

individuals in the model based on their characteristics. Three waves of the HRS include individuals' best estimate of the total cost of their medical care since the previous wave, so I calculate the copay rate faced by each individual as the ratio of out-of-pocket costs to total medical care costs. A copay function is then estimated by OLS by regressing the imputed copay rates on age, sex, health, income, and their interactions. The total insurance premiums paid by each respondent is constructed from the HRS data, and a similar equation is estimated for premiums. The coefficients in the table below are used in the estimation and simulation as the exogenous insurance policies faced by simulated individuals. Fitted copay rates are censored to be between 0.1 and 1, and insurance premiums are censored below by $100. As found by DFJ, most of the coefficients are statistically insignificant, but the copay rate does decline with worse health, reimbursing more generously when the individual is already sick (see Table 4).

4 Estimation of the Structural Model

The model is estimated using the simulated method of moments (SMM), seeking to find parameters that make simulated agents most closely match behavior and outcomes actually observed in the HRS dataset. On the whole, the estimated parameters comport with economic intuition and previous estimates in related models, and most are strongly identified. The estimation method is described in Section 4.1, followed by a discussion of how the moments identify the model parameters in Section 4.2. The estimated model is discussed in Section 4.3.

4.1 Estimation Method

SMM is used to estimate the model because of the diversity of objects to be matched. While the key parameters are identified through differences in health outcomes by income and assets, the model is also concerned with the magnitude of medical expenses faced by individuals, as well as their stock of savings as a buffer against future medical needs. Four basic steps are followed when a set of parameters is tested for fitness. First, the model is solved for optimal behavior at these parameters. Second, the model is simulated using the HRS data to provide initial conditions. Third, the simulated outcomes are compared to the HRS data to generate moments concerning assets, medical expenses, and health. Finally, the many moments are aggregated into a single scalar, which is the object to be optimized in the estimation. Each of these steps are described in more detail below.

Solving and simulating the model: Optimal behavior is solved for each of the one hundred fifty types described, starting from the terminal age of 105 and proceding backwards in two-year periods until age 65 (or the first period observed in the data for that cohort). As described in Section 2.1, (8) and (9) can be successively used to solve for optimal composite consumption and health investment at each point in the state space, with optical medical consumption given by (14). These values are calculated at discrete gridpoints in the three continuous dimensions (b, h, and η), with interpolations used for intermediate values. The range of η values varies greatly by age and health, so this dimension is transformed into the $[0, 1)$ interval of CDF values. Behavioral optimization is conducted by successive grid searches because the choice-payoff function is not necessarily concave. A Monte Carlo simulation of the relevant individuals is conducted; see appendix for details.

Calculating moments: Individuals in the data are categorized in several ways, with "cells" for moments generated by combinations of these categories. First, individuals have an innate sex and cohort (two-year date of birth groups). Second, individuals' income quintile is determined relative to their sex-cohort peers– the method used to sort them into "types". Third, individuals' wealth quintile is determined relative to their sex-cohort-income quintile peers, based on assets in the first observed period. Fourth, individuals' health tertile uses the same reference group, based on health in the first observed period. Finally, individual observations are categorized by the data year and by the age of the individual (in two-year blocks from ages 67-68 to 91-92). These categorizations are combined to generate different classes of cells, described in the appendix.

The six types of moments used concern median assets, median out-of-pocket medical spending, median health, mortality rates, the variance of out-of-pocket expenses, and the correlation between health in successive periods. The moments based on medians measure the proportion of actual observations that are above the simulated median for that cell. The non-median moments (mortality rate, the variance of out-of-pocket medical expenses, and the correlation of health status between periods) use a transformed ratio of the simulated to actual object. All moments have a range of $[-0.5, 0.5]$, putting equal weight on each one when no weighting matrix is applied to the vector of moments. Details on the specific moments employed can be found in the appendix.

Aggregating moments: The vector of all moments is summed using a weighting matrix. The

resulting scalar is the maximand of the estimation:

$$\Xi(\Delta) = -X(\Delta)'WX(\Delta) \tag{15}$$

In this equation, Δ represents the set of parameters to be evaluated for fitness, $X(\Delta)$ is a column vector of all of the moments x calculated at these parameters, and W is the weighting matrix. The weighting matrix that maximizes efficiency is endogenous to the moments themselves, so a standard two-step estimation procedure is used. In the first step, W is a diagonal matrix with elements equal to the cell size for each moment. The optimal weighting matrix implied by the result of the first step (see appendix) is then used as W in the second step to generate the parameter estimates used in the simulations. The object of the estimation is to find the parameters that maximizes $\Xi(\Delta)$– the smallest weighted sum of all of the moments.

4.2 Identification

Thirty-two parameters are estimated by SMM, which can be decomposed into four groups: seven utility parameters, eight shock parameters, eleven health parameters, and six mortality parameters. The utility floor $\underline{u}$ is calibrated from DFJ's estimation rather than estimated jointly. In a complex structural model, a change in any parameter will shift nearly all of the moments; the identification arguments presented below focus on the moments most directly related to each parameter and ignore secondary effects.

The intertemporal discount factor δ and the coefficient of relative risk aversion for composite consumption ρ are identified through the median assets moments for the income-wealth-year cells. Larger values of δ mean that individuals put more weight on future utility, inducing them to save more assets. Risk aversion has a similar effect on savings, but differentially by income and age. Older and higher income individuals face more risk in their level of composite consumption because they have greater medical needs risk and because they are less protected by the utility floor– their higher expected consumption means they have "further to fall" before being caught by the social safety net. As discussed by Attanasio and Low (2004), the growth rate of consumption should be positively correlated with variance in consumption, so older and richer individuals accelerate their asset depletion more rapidly. Larger values of ρ amplify this effect, and thus risk aversion with respect to composite consumption is identified separately from the discount factor.

The curvature and scale of the bequest motive (ω_0 and ω_1 respectively) are also identified through

the median assets moments for the wealth-income-age cells, in particular the variation in wealth. For individuals with sufficiently high assets, the bequest motive is inoperative: the marginal utility of consumption they would obtain in the absence of any bequest motive is large compared to the marginal utility of their potential bequest, so they are unwilling to give up much consumption for the sake of a larger bequest. Thus we expect individuals in lower wealth quintiles to have larger deviations from the asset profiles that would occur without a bequest motive. With five wealth quintiles within each income quintile, I am able to identify both the curvature and the scale of the bequest motive.

The coefficient of relative risk aversion for medical consumption, ν, is identified through the median medical spending moments for the income-wealth cells. Recall that (14) gives the optimal amount of medical consumption as a function of composite consumption. The rate at which median medical spending increases across wealth quintiles thus identifies the ratio ρ/ν; as ρ is identified by the estimation, so is ν. The marginal utility shifter for health α_1 is identified through the median asset moments for the sex-income-health-year cells. If the marginal utility of composite consumption changes with health, then there will be systematic deviations from the asset profiles with respect to health: lower marginal utility of consumption should induce slower asset depletion. Note that the bequest motive can also generate this effect, as the larger probability of death with poor health induces individuals to save more; α_1 is still identified, as it does not have a corresponding systematic effect on the income-wealth-year asset profiles.

The four parameters that chiefly govern the quantity of health investments purchased are identified through the median health moments for the income-wealth-age cells and the sex-income-health-year cells. These parameters include the scale and curvature of the efficacy of health investments (γ_{m1} and γ_{m0}), the effect of illness on the efficacy of health investments γ_{m2}, and the utility shifter for life ς_0. The utility level shifter ς_0 is identified by the wealth-income groups that do not purchase any health investment and thus do not deviate from the expected rate of health decline (given the other γ parameters). For individuals with very little assets or income, consumption levels are so low that their expected utility within any given period is negative. These individuals will not seek to extend their life through health investments, as this would decrease their expected lifetime utility. The level of wealth and income where the motive to extend life becomes operative thus identifies ς_0 by determining the level of expected consumption at which an individual would be indifferent to health investments. Identification is satisfied so long as at least one adjacent income-wealth group has an identical rate of health decline to the absolute poorest group (first income quintile, first

wealth quintile).

The efficacy and curvature of health investments are identified by the extent of the deviations from the expected path of health decline across the upper levels of wealth and income. The larger the value of γ_{m1}, the slower median health will decline for groups with high income and/or assets as individuals in these groups will invest more in their health. The curvature γ_{m0} is identified by the degree to which the marginal change in health decline tapers off at the very highest income and wealth groups. Holding constant the slope of the health investment production function when there is no health investment (i.e. $\exp(\gamma_{m1} + \gamma_{m2}(1-h) - \gamma_{m0})$), lower values of γ_{m0} mean that the marginal returns to health investments decline more rapidly. Reduced form analysis reveals that the rate of health decline is nearly identical among the richest wealth-income groups, identifying the curvature of the health investment production function. The effect of illness on the efficacy of health investments (γ_{m2}) is identified through the median health moments for the health-sex-income-year cells.

Identification of the remaining parameters is relatively straightforward and does not warrant much discussion. The scale parameters for the distribution of medical needs (β_0, β_s, β_{a1}, β_{a2}, β_{h1}, and β_{h2}) are identified by the median out-of-pocket spending moments for the age-sex cells and the health thirds cells. The shape parameters for the medical needs distribution (β_{k0} and β_{k1}) are identified by the variance of out-of-pocket expense moments by health thirds. The remaining parameters of the health evolution distribution (γ_0, γ_s, γ_{a1}, γ_{a2}, γ_{h1}, and γ_{h2}) are identified by the median health moments for the sex-income-health-year cells and the sex-cohort-year cells. The parameters governing the magnitude of health shocks (σ_0 and σ_1) are identified by the two health correlation moments. Finally, the mortality parameters (θ_0, θ_s, θ_{a1}, θ_{a2}, θ_{h1}, and θ_{h2}) are identified by the mortality rate moments for the same cells.

4.3 Estimated Model

Estimates of the model parameters can be found in Table 1 and generally align with previous models' findings and intuition about the parameters. The estimated value of δ corresponds to an annual discount factor of 0.968, similar to previous estimates. Relative risk aversion for composite consumption ($\rho = 2.06$) and for medical consumption ($\nu = 3.28$) are both very similar to DFJ's estimates in the analogous specification. In further concordance with DFJ, the estimated bequest motive is fairly small and is not a significant motivator of savings for the retired population. Figure

4 shows the model's fit of income-conditional median asset profiles; the model fits the data rather well, but somewhat overestimates the rate of asset decline for the top income quintile.

Interpretation of the value of life ($\varsigma_0 = 0.293$) and the marginal utility shifter ($\alpha_1 = -0.46$) is somewhat difficult, as with $\rho > 1$ and endogenous lifespans, the marginal utility effect of α_1 is confounded with its level effect. The negative value indicates that the marginal value of consumption is lower with better health, but the level of utility is higher. The estimation might produce this value not because of variation in asset depletion across health, but because health brings utility benefits beyond lower medical needs. These parameters indicate that the critical consumption value at which an individual gets utility exactly equal to the value of being dead is about \$9,900 (annual) for an individual in rather good health ($h = 0.8$) and \$13,700 for an individual in fairly poor health ($h = 0.2$), assuming a small medical needs shock (see Figure 2). While individuals with expected consumption levels below these will experience negative utility, they still may seek to maintain or improve their health to avoid the higher medical costs and lower utility from consumption associated with illness. DFJ's estimate of the analogue of α_1 as approximately -0.2 is comparable to my own, as they use a binary specification for health rather than a continuous measure.

The estimates of the medical needs shock distribution β largely comport with expectations. The base values of the scope parameter ($\beta_0 = -21.9$) and the shape parameter ($\beta_{k0} = 0.11$) indicate that the distribution generates many very small values, with its mean driven by an extreme upper tail. This accurately describes the observed distribution of out-of-pocket medical expenses, and the simulated distribution is a good fit to the data as seen in Figure 3. The distribution of medical needs is lower for men, whose medical expenses are driven by their lower average health and greater financial resources. As expected, medical needs rise steeply with age and illness; the age effect is nearly linear ($\beta_{a1} = 0.307$, $\beta_{a2} = 0.0006$), while the health effect is driven by the quadratic term for illness ($\beta_{h1} = 2.57$, $\beta_{h2} = 9.36$). The health-conditional simulated distributions of out-of-pocket expenses also match the data well.

The health transition parameters γ are likewise unsurprising. The mean of the subsequent health state is roughly equal to the current health state, as the effect of current on health on future health is mostly linear ($\gamma_0 = 0.026$, $\gamma_{h1} = 0.82$, $\gamma_{h2} = 0.16$). The rate of health decline accelerates with age ($\gamma_{a1} = -0.0025$), and men's health declines faster on average than women's ($\gamma_s = -0.005$). Health evolution shocks have a standard deviation of about 10% of the range of health for healthy individuals, and about 14% for very sick individuals; that is, health is more stable for individuals in good health. The parameters governing the efficacy of health investments indicate that while the

initial "health dividend" is quite high (as the difference between γ_{m1} and γ_{m0} is large), the returns to health investment taper off quite rapidly due to the low value of γ_{m0}. The health production function implied by these parameters can be seen in Figure 8: the first few hundred dollars of health investment lead to large gains in health, but spending beyond about $1000 yields very small marginal gains. The mortality parameters θ all align with intuition: old age, poor health, and being male all contribute to a higher likelihood of death above and beyond that predicted by the base health transition parameters. As seen in Figure 9, the simulated distribution of longevity matches the data fairly well, but has slightly less variance than it should.

The estimated model matches the median health profiles across income quintiles very well, as seen in Figures 6 and 7. The model matches the second and third income quintiles nearly perfectly, while showing a slightly better profile for the poorest quintile and faster rate of health decline for the fourth and fifth income quintiles. There is a tension in the estimation between matching medical spending distributions and matching median health profiles: "nearby" parameter sets that would boost health investment of the highest income quintile to more accurately match their health profile would result in simulated medical spending distributions that are poor fit to the data. The model's estimates are a compromise between the two objectives.

5 Counterfactual Policy

With the estimated model in hand, the effects of a hypothetical subsidy on preventive care can now be evaluated. Relative to the baseline scenario without the subsidy, individuals targeted by the subsidy experience somewhat extended lives and a reduction in out-of-pocket medical expenses. The subsidy is not cost-saving for the government for any sub-population, but the costs are relatively small when spread over among all retirees. Section 5.1 describes the procedure used in the counterfactual simulation and the metrics used to evaluate the subsidy, while Section 5.2 analyzes the effects of the subsidy based on these metrics.

5.1 Counterfactual Procedure

I gauge the effects of two hypothetical policies that subsidize the health investment good: a "universal" subsidy and a "targeted" subsidy. The universal subsidy reimburses 75% of out-of-pocket costs from health investment for all individuals without limit, while the targeted subsidy adjusts its

generosity based on health and income. The new budget constraint, replacing (7), is:

$$b_{it+1} = R_t \left(b_{it} + y_{it} + w_{it} - z_{it} - c_{it} - q_{it} p_t (\mu_{it} + \kappa_{it}) + s_{it} \right) \geq 0. \tag{16}$$

The universal and targeted policies are respectively:

$$s_{it} = (1 - q_{it}) p_{it} \kappa_{it} \cdot 0.75. \tag{17}$$

$$s_{it} = \min((1 - q_{it}) p_{it} \kappa_{it} \min(\max(4.75 - 2y_{it}, 0), 0.9), 0.05) \quad \text{if} \quad h_{it} \geq 0.5. \tag{18}$$

The targeted subsidy is designed to optimize the cost efficiency of the policy, achieving the greatest gains at low cost. Healthier individuals expect to live longer and thus reap greater benefits; the targeted subsidy is thus only available to those in sufficiently good health. Moreover, this restriction can be interpreted as subsidizing preventive care rather than curative care, both forms of health investment. The targeted policy is also means-tested by decreasing its generosity as income increases. Estimation reveals that higher income individuals already make significant investments in their health, so means-testing alleviates individuals being subsidized for purchases they would have made anyway. I cap the subsidy at \$500 per two-year period per individual so that those with very high assets do not exploit the subsidy for very little health gain. The targeted subsidy effectively only applies to individuals in the first and second income quintiles, as it is completely phased out for someone with annual income of \$13,000 (\$26,000 in model terms).

Initial conditions for the counterfactual simulations are given by the most recent wave of the HRS data (2010). The sample used includes individuals born in or before 1945 who were unmarried and retired as of the 2010 data collection. These criteria admit three (two-year) cohorts that were not included in the sample used for estimation, allowing an analysis of the effects of the subsidy on younger retirees who have a longer expected remaining lifetime. The population is not refreshed with new retirees as the simulation progresses because the model does not apply to the working or married population, nor can it predict when these individuals will retire or become widowed. Each individual's initial conditions are replicated five hundred times so that the aggregate simulated values are not sensitive to individual shocks. The model with the subsidy is simulated from 2010 until all individuals have died; a simulation without the subsidy is also conducted for a baseline comparison.

Several objects of interest are calculated to evaluate the effects of the counterfactual subsidy.

First, the simulation tracks the age at death for each individual, to find the effect of the subsidy on longevity. Second, both total and out-of-pocket medical expenses are recorded over the lifespan of each simulated agent; the lifetime sum of these values is calculated, discounted by the interest rate. Third, the simulation tracks government expenditures on each individual in the form of welfare payments (for individuals who cannot achieve the utility floor), subsidy payments, and Medicare reimbursements (assumed to be 60% of insurance payouts).[5] The discounted sum of these expenses, differenced between the counterfactual and the baseline and averaged across copies of an individual, yields the expected lifetime cost of the policy for each individual. Finally, I calculate each individual's willingness to pay for the subsidy as the compensating variation between the counterfactual and the baseline– the asset loss that leaves an individual indifferent between the baseline with no loss and the counterfactual policy with the loss. These measures allow consideration of both the costs and benefits of the subsidy policy. To gauge the lifetime effect of the subsidy, I focus on retirees of ages 65-70 in 2010; the costs and benefits calculated for older retirees does not reflect the ongoing effect of the policy for future generations.

5.2 Effects of a Subsidy on Preventive Care

Universal Subsidy: The universal subsidy extends the life of retirees by an average of 0.53 months, but incurs an additional $38,425 in government expenditures per capita over the life of a retiree. Table 5 decomposes the longevity gains by income quintile and health, and Table 6 decomposes them by income and wealth quintiles. Lifetimes seem to be most extended for individuals in the middle range of income and wealth: the very poorest individuals do not avail themselves of the subsidy, while it is less effective for very wealthy individuals who already make significant investments in health without the subsidy. The gains are also generally larger for individuals already in good health.

The decomposition of the costs of the universal subsidy are found in Tables 7 and 8. Intuitively, wealthy individuals– both in income and assets– account for the very large cost of the subsidy, while the poorest individuals incur nearly no public expense at all. In combination with their existing insurance, the rich pay very little out-of-pocket for health investment when offered the universal subsidy. Their high levels of consumption make additional periods of life very desirable, so they purchase enormous quantities of health investment. As seen in Figure 8, the returns to health

[5] In a future version of this paper, the portion of medical expenses paid by Medicare will be estimated rather than use an assumed constant rate. This calculation requires Medicare data linked to the HRS that has not yet been acquired.

investment decrease rapidly, so the tens of thousands of dollars spent on medical care amount to no greater effect than the more reasonable purchases of middle income individuals.

The absurdly large costs for the wealthiest retirees are a product of the model's parametric form and should not be taken too seriously– it is likely impossible to purchase $25,000 worth of medically relevant health investment every single year. It is nonetheless reasonable to conclude that a subsidy on health investment is particularly inefficient when offered without limitation to wealthy retirees. Analysis of the universal policy motivates the restrictions placed on the subsidy for the targeted policy. The targeted subsidy phases out for individuals with moderate incomes, is only available to reasonably healthy retirees, and is capped at $500 biannually. The last provision is influenced by the severe curvature of the health production function: nearly all of the health gains are attained in the first few hundred dollars spent on health investment. The subsidy cap prevents wasteful spending by individuals with high assets, increasing the cost efficiency.

Targeted Subsidy: The targeted subsidy is effective in extending the lifespans of some individuals and marginally reducing out-of-pocket medical expenses, but it is not cost-saving for the government, nor does it reduce total medical spending. Because of the income restriction placed on the subsidy, it is effectively only available to individuals in the first and second income quintile. The requirement that recipients of the subsidy already be in sufficiently good health causes the effects of the subsidy to be concentrated among those who meet this requirement at the outset of the policy; those below the threshold are less likely to be able to prevail themselves of the subsidy. I thus label retirees in the bottom two income quintiles whose health is greater than 0.5 in 2010 as the "target population". The targeted subsidy increases life expectancy among retirees age 65-70 in the target population by 0.76 months, at a lifetime public cost of $760 per capita ($167 when spread over the population of young retirees).

Table 9 presents the change in life expectancy by income and health, while Table 10 decomposes these changes by income and wealth. The increase in longevity is mostly achieved by retirees who are already in good health. Moreover, the very poorest individuals– those in the lower wealth quintiles of the bottom income quintile– do not see much benefit from the subsidy. The model predicts that the very poorest individuals will not invest in their health even when it is offered for free because their consumption is so low as not to motivate a desire for longer life. Further, even individuals whose consumption does not reach the critical level prefer better health because it improves their distribution of medical needs and thus out-of-pocket costs; for the poorest individuals, this does not

bind because they are so often at the utility floor.

Notably, the effect of wealth on the change in longevity varies between the first and second income quintile. Among the lowest income group, retirees with low assets see very little gain from the subsidy because they largely do not take advantage of it. For the poorest individuals, expected utility each period is so low that they would not invest in their health even if the investment were free. Thus the effect of the subsidy is greater for individuals in the bottom income quintile whose assets are sufficient to raise their consumption (and expected utility per period) beyond the critical level. In contrast, the longevity improvement among retirees in the second income quintile decreases with greater wealth. The income of the second quintile is sufficient to meet the critical level of consumption even at very low levels of assets, but those in higher wealth quintiles already make significant investments in their health in the absence of the subsidy. Thus the gains from the subsidy are greatest for those who do not already invest in their health, but are willing to if the price is right.

As seen in Tables 11 and 12, the targeted policy is not cost-saving for any subgroup. As expected, healthy individuals are the primary drivers of the cost of the subsidy, as sicker retirees must draw into improved health before the subsidy is available to them. The large cost of the policy among the healthiest individuals is a result of these retirees living longer and thus being subsidized more often as well as the coincidence of high health with the upper wealth quintiles. Even with the subsidy capped at $500 each period, the top wealth quintile is still the major driver of costs. Critically, the potential cost savings from the subsidy (due to reduced Medicare payments from lower purchases of medical consumption, as well as lower welfare payments to individuals who cannot achieve the utility floor) do not occur. In fact, the targeted subsidy *increases* purchases of medical consumption of the remaining lifetime of young retirees. Figures 10 and 11 show the distributions of targeted retirees' valuations of the subsidy and expected government costs incurred. Though the subsidy is cost-saving for about 14% of the targeted population, the scale of the costs generally outpaces individuals' willingness to pay for it.

Rather than have offsetting effects on public expenditures, the government ultimately pays for the subsidy twice. Rather than simply increasing the amount of health investment purchased, individuals use most of the savings from the subsidy to increase their composite consumption and medical consumption. As the government partially bears the cost of medical consumption spending through Medicare, they pay for the subsidy both directly and indirectly. As a secondary effect, the extended lifetimes of some subsidy recipients create an additional period of Medicare payments for

these individuals, further increasing costs. The subsidy is only cost-saving for the government in states of the world in which the individual has a very bad medical needs shock; these large shocks are insufficiently common to offset the more likely outcome that individuals purchase more medical consumption due to the income effect of the subsidy. The direct cost of the subsidy accounts for $340 per targeted individual, while the remaining $420 is indirect spending.

Lifetime medical expenditures among the targeted population rise by an average of $979, while out-of-pocket costs fall marginally, by about $60. Individuals thus take the benefit of the subsidy largely in the form of increased consumption and medical consumption, with very little effect on savings. Though the government bears most of the increase in medical expenditures, reimbursements by private insurers would increase by an average of $279 over the lifetime of a targeted individual. While the model did not dynamically price private insurance policies (by using a zero profit condition, e.g.), the policy could cause a small increase in health insurance premiums if this effect were accounted for.

The critical identifying assumption of my estimation— that differences in rates of health decline across income and wealth are entirely attributable to investments in health— should tend to amplify the estimated efficacy of health investment and thus bias the counterfactual simulation to reveal greater benefits for a subsidy on preventive care. Even with this bias in favor of the policy, I still do not find it to be cost saving nor able to increase longevity more than marginally. It is thus likely that more complex models that allow for multiple pathways between income and health (such as homogeneity of preferences or non-medical health behaviors) would yield even less promising estimates of the effects of the subsidy.

Sensitivity Analysis: To better understand how the parameters of the model influence the effects of the subsidy policy, I conduct a sensitivity analysis with respect to three key parameters: the value of life ς_0, the curvature of the health production function γ_{m0}, and the efficacy of health investments γ_{m1}. For each of the parameters, I vary its value significantly above and below the estimated value and re-simulate the targeted policy counterfactual, tracking several measures of the costs and benefits of the policy: the mean change in life expectancy among the targeted population, the mean change in total lifetime medical expenses, the distribution of valuations of the subsidy, and the distribution of net government costs of the subsidy.

Figure 12 plots the results of the sensitivity analysis with respect to ς_0. Higher levels of the value of life are associated with greater medical expenses, valuations, and public costs of the subsidy,

as retirees place greater emphasis on extending their lives and thus take greater advantage of the subsidy. However, higher ς_0 does not necessarily mean that the policy is more effective in actually improving longevity. In the upper left panel of Figure 12, the relationship between mean change in longevity and ς_0 is not monotonic, with the model's estimate near the top of the curve. If ς_0 were higher than estimated, targeted retirees would make further investments in their health even without the subsidy; the additional health investment induced by the subsidy thus occurs further along the health production function, where the marginal gains are smaller. On the other hand, at values of ς_0 less than about 0.1, none of the targeted retirees purchase health investments with or without the subsidy, so it has no cost or benefit.

Results of the sensitivity analysis with respect to γ_{m0} are shown in Figure 13. The estimation poorly identified the curvature of the health production function, so the range tested in this analysis is quite large. At large values of γ_{m0}, the initial slope of the health production function is too shallow to induce targeted individuals to purchase any health investment, and thus the subsidy is completely ineffective. As γ_{m0} becomes very negative, the initial slope of the health production function becomes steep even while the marginal returns to health investment dwindle faster. On net, lower values of the curvature parameter are associated with greater valuations and government costs, as poorer individuals find at least some health investment worth the cost. In contrast to the previous analysis, the growth of the mean valuation of the subsidy seems to outpace the growth of the mean public costs.

Finally, Figure 14 shows the sensitivity analysis with respect to γ_{m1}, the efficacy of health investments (level of the health production function). Intuitively, as health investment produces more health, the subsidy's effectiveness in extending longevity among the targeted population also increases rapidly. At the top of the tested range of γ_{m1}, the subsidy extends lifetimes by about 9 months on average, about twelve times the improvement in the estimated model. However, these very high levels of γ_{m1} are a poor fit for the data, as lifespans would be significantly longer across the population even without the subsidy, while out-of-pocket medical spending would be much higher. The size of the standard error is likely attributable to the relatively flat region of sensitivity curves: in a local region around the estimated value, outcomes do not vary drastically. As before, the greater benefits of the subsidy are associated with higher costs in terms of both total medical spending and additional payments by the government. As with ς_0, there seems to be a limit to the growth of the longevity gains from the subsidy; at very high values of γ_{m1}, the benefits taper off as even poor individuals invest heavily in their health without the subsidy.

6 Conclusion

This paper posits a structural model of the retirement life cycle that includes two medical care goods. This structure allows medical expenditures to be both highly variable (through medical needs shocks that determine demand for the consumption-type medical care good) and to influence the future health state (through purchases of the investment-type medical care good) without tying shocks to health to the variance of medical expenses. The model is estimated using panel data from the Health and Retirement Study, simultaneously matching conditional profiles of asset holdings, out-of-pocket medical expenses, health, and mortality. The estimated model does a fairly good job of matching each of these objects across the spectrum of income and wealth. Estimation reveals that initial investments in health are particularly important to the subsequent health state, but the marginal returns to health investment drop off rapidly.

Counterfactual simulation of a subsidy on the investment-type medical care good shows that such a policy can somewhat extend the lives of retirees, but would come at great public expense unless restrictions are placed on the subsidy. A narrowly tailored subsidy targeted at lower income retirees in good health (subsidizing preventive rather than curative care) that caps the benefit is able to produce similar gains in longevity at a small fraction of the cost of the universal subsidy. In theory, a subsidy on preventive care could be cost-saving for the government if the resulting improvement in health sufficiently reduces future medical care spending so as to offset the upfront cost of the subsidy. However, the counterfactual simulation reveals that the subsidy is not cost-saving for any subgroup of income, wealth, or health. Indeed, retirees use the savings from the subsidy to purchase more consumption and consumption-type medical care. The government thus pays for the subsidy twice, as it also bears a large fraction of medical care costs through Medicare.

The model accounts only for single, retired individuals and is not necessarily predictive of behavior for coupled retirees or the working population. Importantly, the estimation and counterfactuals do not account for the behavior of workers in anticipation of retiring with a subsidy on health investments in place. Individuals might seek to better preserve their health before retirement so as to take advantage of the subsidy, reducing future Medicare reimbursements. Health investments through preventive care might also be more effective for the younger individuals (and there is a longer period for health to accrue from the investment), so that a subsidy might be more cost-efficient when applied to the working population. Because the model fits the data reasonably well, it is worthwhile to extend the model to include working life and the joint optimization problem of spouses. Moreover,

non-medical health behaviors (such as smoking, drinking, and exercise) can be incorporated into the model to further explain differences in health depreciation across individuals.

References

Ameriks, J., Caplin, A., Laufer, S., and Nieuwerburgh, S. V. (2011). "The Joy of Giving or Assisted Living? Using Strategic Surveys to Separate Public Care Aversion from Bequest Motives." *Journal of Finance*, 66(2): 519–561.

Arcidiacono, P., Sieg, H., and Sloan, F. (2007). "Living Rationally Under the Volcano? An Empirical Analysis of Heavy Drinking and Smoking." *International Economic Review*, 48(1): 37–65.

Attanasio, O. and Low, H. (2004). "Estimating Euler Equations." *Review of Economic Dynamics*, 7: 405–435.

Bajari, P., Hong, H., Khwaja, A., and Marsh, C. (2006). "Moral Hazard, Adverse Selection and Health Expenditures: A Semiparametric Analysis." *NBER Working Paper*, (12445).

Blau, D. M. and Gilleskie, D. B. (2006). "Health Insurance and Retirement of Married Couples." *Journal of Applied Econometrics*, 21(7): 935–953.

Blau, D. M. and Gilleskie, D. B. (2008). "The Role of Retiree Health Insurance in the Employment Behavior of Older Men." *International Economic Review*, 49(2): 475–514.

Bound, J., Stinebrickner, T. R., and Waidman, T. A. (2010). "Health, economic resources and the work decisions of older men." *Journal of Econometrics*, 156(1): 106–129.

Carroll, C. D. (1997). "Buffer Stock Saving and the Life Cycle/Permanent Income Hypothesis." *Quarterly Journal of Economics*, 112(1): 1–56.

Carroll, C. D. and Samwick, A. A. (1997). "The Nature of Precautionary Wealth." *Journal of Monetary Economics*, 40(1): 41–71.

Case, A. and Deaton, A. (2003). "Broken Down by Work and Sex: How Our Health Declines." *NBER Working Paper*, 9821.

Cohen, J. T., Neumann, P. J., and Weinstein, M. C. (2008). "Does Preventive Care Save Money? Health Economics and the Presidential Candidates." *New England Journal of Medicine*, 358(7): 661–663.

Conley, D. and Thompson, J. A. (2011). "Health Shocks, Insurance Status, and Net Worth: Intra- and Inter-Generational Effects." *NBER Working Paper*, 16857.

Currie, J. (2008). "Healthy, Wealthy, and Wise: Socioeconomic Status, Poor Health in Childhood, and Human." *NBER Working Paper*, (13987).

De Nardi, M., French, E., and Jones, J. B. (2010). "Why Do the Elderly Save? The Role of Medical Expenses." *Journal of Political Economy*, 118(1): 37–75.

Deaton, A. (2002). "Policy Implications Of The Gradient Of Health And Wealth." *Health Affairs*, 21(2): 13–30.

Deaton, A. (2003). "Health, Inequality, and Economic Development." *Journal of Economic Literature*, 41: 113–138.

DePreux, L. B. (2011). "Anticipatory Moral Hazard and the Effect of Medicare on Prevention." *Health Economics*, 20: 1056–1072.

Finkelstein, A. (2002). "When Can Partial Public Insurance Produce Pareto Improvements?" *NBER Working Paper*, (9035).

Finkelstein, A. and Poterba, J. (2006). "Testing for Adverse Selection with "Unused Observables"." *NBER Working Paper*, (12112).

Finkelstein, A., Luttmer, E. F., and Notowidigo, M. J. (2008). "What Good Is Wealth Without Health? The Effect of Health on the Marginal Utility of Consumption." *NBER Working Paper*, (14089).

French, E. (2005). "The Effects of Health, Wealth, and Wages on Labor Supply and Retirement Behavior." *Review of Economic Studies*, 72: 395–427.

French, E. and Jones, J. B. (2011). "The Effects of Health Insurance and Self-Insurance on Retirement Behavior." *Econometrica*, 79(3): 693–732.

Fries, J. F., Koop, C. E., Beadle, C. E., Cooper, P. P., England, M. J., Greaves, R. F., Sokolov, J. J., and Wright, D. (1993). "Reducing Health Care Costs By Reducing the Need and Demand for Medical Services." *New England Journal of Medicine*, 329(5): 321–325.

Gilleskie, D. B. (1998). "A Dynamic Stochastic Model of Medical Care Use and Work Absence." *Econometrica*, 66: 1–45.

Grossman, M. (1972). "On the Concept of Health Capital and the Demand for Health." *Journal of Political Economy*, 80(2): 223–255.

Hall, R. and Jones, C. (2007). "The Value of Life and the Rise in Health Spending." *NBER Working Paper*, 10737.

Hugonnier, J., Pelgrin, F., and St-Amour, P. (2012). "Health and (Other) Asset Holdings." *Review of Economic Studies*, 9(4).

Kenkel, D. (). *Handbook of Health Economics*, vol. 1, chap. 31, pp. 1675–1720. Elsevier.

Khwaja, A. (2010). "Estimating Willingness to Pay for Medicare Using a Dynamic Life-Cycle Model of Demand for Health Insurance." *Journal of Econometrics*, 156: 130–147.

Lockwood, L. (2012). "Incidental Bequests: Bequest Motives and the Choice to Self-Insure Late-Life Risks." *NBER Working Paper*.

Ozkan, S. (2011). "Income Inequality and Health Care Expenditures over the Life Cycle." *Working paper*.

Åke Blomqvist (1991). "The Doctor as Double Agent: Information Asymmetry, Health Insurance, and Medical Care." *Journal of Health Economics*, 10: 411–432.

Russell, L. B. (1993). "The Role of Prevention in Health Reform." *New England Journal of Medicine*, 329(5): 352–354.

Russell, L. B. (2007). "Prevention's Potential For Slowing the Growth of Medical Spending." *National Coalition on Health Care*.

Russell, L. B. (2009). "Preventing Chronic Disease: An Important Investment, But Don't Count On Cost Savings." *Health Affairs*, 28(1): 42–45.

Thorpe, K. E., Florence, C. S., and Joski, P. (2004). "Which Medical Conditions Account For The Rise In Health Care Spending?" *Health Affairs*, W4.

van der Klauw, W. and Wolpin, K. I. (Social Security and the Retirement and Savings Behavior of Low Income Households). "2008." *Journal of Econometrics*, 145(1-2): 21–42.

Vera-Hernàndez, M. (2003). "Structural Estimation of A Principal-Agent Model: Moral Hazard in Medical Insurance." *RAND Journal of Economics*, 34(4): 670–693.

Weisbrod, B. A. (1991). "The Health Care Quadrilemma: An Essay on Technological Change, Insurance, Quality of Care, and Cost Containment." *Journal of Economic Literature*, 29: 523–552.

Yang, Z., Gilleskie, D. B., and Norton, E. (2009). "Health Insurance, Medical Care, and Health Outcomes: A Model of Elderly Health Dynamics." *Journal of Human Resources*, 44(1): 47–114.

Yogo, M. (2009). "Portfolio Choice in Retirement: Health Risk and the Demand for Annuities, Housing, and Risky Assets." *NBER Working Paper*, (15307).

Table 1: Parameters Estimated by the Simulated Method of Moments

Parameter	Value	Std Error	Description
δ	0.937	(0.0036)	Intertemporal discount factor
ρ	2.062	(0.0355)	Coefficient of risk aversion for composite consumption
ν	3.283	(0.0646)	Coefficient of risk aversion for medical consumption
ς_0	0.293	(0.0081)	Utility level shifter; value of living
α_1	-0.459	(0.0192)	Change in marginal utility with health
$\underline{u}$	-4.0	(N/A)	Utility floor (not estimated)
ω_0	1.285	(0.126)	Curvature of bequest motive
ω_1	4.184	(0.291)	Intensity of bequest motive
β_0	-21.96	(0.0284)	Base value of medical needs scope parameter
β_s	-2.178	(0.201)	Change in medical needs scope for males
β_{a1}	0.307	(0.0093)	Change in medical needs scope with age
β_{a2}	0.00064	(0.00011)	Change in medical needs scope with age squared
β_{h1}	2.575	(0.202)	Change in medical needs scope with sickness
β_{h2}	9.356	(0.254)	Change in medical needs scope with sickness squared
β_{k0}	0.108	(0.0007)	Base value of medical needs shape parameter
β_{k1}	0.011	(0.0017)	Change in medical needs shape with health
γ_0	0.0264	(0.0014)	Base health transition level
γ_s	-0.00495	(0.0005)	Change in health transition for males
γ_{a1}	-0.00246	(0.00012)	Change in health transition with age
γ_{a2}	-0.000006	(0.000005)	Change in health transition with age squared
γ_{h1}	0.820	(0.0006)	Change in health transition with health
γ_{h2}	0.159	(0.0021)	Change in health transition with age squared
γ_{m0}	-12.48	(2.33)	Curvature of health investment production
γ_{m1}	-7.58	(0.0971)	Base efficacy of health investment
γ_{m2}	-0.0155	(0.011)	Change in efficacy of health investment with sickness
σ_0	0.0985	(0.0006)	Base standard deviation of health transition
σ_1	0.0417	(0.0016)	Change in health transition s.d. with sickness
θ_0	-2.687	(0.0015)	Base mortality probit level
θ_s	0.452	(0.0385)	Change in mortality for males
θ_{a1}	0.0170	(0.00007)	Change in mortality with age
θ_{a2}	0.00139	(0.000001)	Change in mortality with age squared
θ_{h1}	1.471	(0.0039)	Change in mortality with sickness
θ_{h2}	0.608	(0.0020)	Change in mortality with sickness squared

Note: All estimated parameters are significant at 1% level or better except γ_{a2} and γ_{m2}, which are not significant.

Table 2: Ordered Probit of Subjective Health on Objective Health

Variable description	Coefficient	Std Err
Male	-.248***	.0066
High blood pressure	-.222***	.0065
Very high blood pressure	-.403***	.0234
Has diabetes	-.390***	.0086
Complications from diabetes	-.343***	.0289
Ever had cancer	-.201***	.0092
Lung disease, etc	-.360***	.0108
Heart conditions	-.296***	.0077
Has ever had a stroke	-.109***	.0163
Ongoing problems from stroke	-.185***	.0235
Psychological problem	-.214***	.0089
Memory problem	-.132***	.0220
Has arthritis	-.075***	.0069
Fallen in past month at all	.042***	.0113
Number of times fallen	-.0035*	.0021
Was hurt in at least one fall	-.00051	.0157
No. of days with lost urine (past month)	-.00074*	.0004
Usually in at least mild pain	-.255***	.0116
Usually in at least moderate pain	-.107***	.0132
Usually in very bad pain	-.284***	.0151
Has depression	-.192***	.0154
No. of days spent in bed (past month)	-.024***	.0012
Difficulty jogging	-.245***	.0094
Difficulty walking a few blocks	-.295***	.0010
Difficulty walking one block	-.119***	.0124
Difficulty sitting down on chair	-.094***	.0087
Difficulty standing up from chair	-.038***	.0080
Difficulty climbing several flights of stairs	-.252***	.0080
Difficulty climbing one flight of stairs	-.129***	.0107
Difficulty stooping to pick up object	-.045***	.0079
Difficulty reaching outward with arms	-.136***	.0095
Difficulty pushing a chair across room	-.170***	.0090
Difficulty carrying a bag of groceries	-.161***	.0097
Difficulty picking up a dime	-.050***	.0129
Difficulty using a map	-.272***	.0081
Need help cooking meals for self	-.102***	.0181
Need help shopping for groceries	-.123***	.0154
Need help using the phone	-.084***	.0185
Need help managing prescriptions	-.0285	.0201
Need help managing personal money	.029*	.01691
Received a flu shot this year	.020**	.0082
Had cholesterol tested this year	-.021**	.0094
Prostate exam / pap smear / mammogram	.085***	.0091
Cutoff 1	-3.260	.0128
Cutoff 2	-2.080	.0112
Cutoff 3	-.953	.0103
Cutoff 4	.219	.0101

Table 3: Income and Wealth Summary Statistics (Sample)

	Income	Assets
Mean	$20,815	$195,274
Std Dev	$36,275	$365,919
1st P	$458	$0
5th P	$5,096	$0
10th P	$6,264	$0
25th P	$8645	$9,360
50th P	$13,727	$78,601
75th P	$23,071	$222,000
90th P	$39,299	$497,904
95th P	$53,947	$791,449
99th P	$117,650	$1,693,375

Table 4: Estimates of Premiums and Copay Rates

	Premiums		Copay Rate	
Variable	Coefficient	t-stat	Coefficient	t-stat
Constant	-.0058	-0.39	.0436	1.56
Health	.0803	1.23	.4325***	3.81
Health squared	.0130	0.20	-.1505	-1.36
Age (minus 65)	.0036	1.59	.0034	0.81
Age squared	-.00008	-1.04	-.000013	-0.09
Male	-.0088***	-2.97	-.0409***	-9.61
Income	.0216***	5.90	.0337***	4.80
Income squared	-.0001***	-4.15	-.0012***	-4.25
Income cubed	6.73e-09***	7.15	6.06e-07	1.55
Health * age	.0039	0.38	-.0067	-0.38
Health squared * age	-.0068	-0.64	.0032	0.18
Health * age squared	.000034	0.10	4.38e-06	0.01
Heath sq * age sq	-.000024	-0.06	-.00012	-0.17
Health * income	-.0324***	-2.66	-.0539**	-2.51
Health squared * income	.0133	1.38	.0268	1.59
Health * income squared	.00007	0.87	.0024***	3.28
Health sq * income sq	.000032	0.54	-.00136***	-2.91

Table 5: Longevity Change in Months from Universal Subsidy by Income and Health

Income		Range of Health h			
Quintile	All	$(0, 0.25]$	$(0.25, 0.5]$	$(0.5, 0.75]$	$(0.75, 1]$
Bottom	0.26	0.02	0.08	0.35	0.88
Second	0.68	0.22	0.54	0.92	0.82
Third	0.58	0.31	0.60	0.64	0.51
Fourth	0.61	0.32	0.52	0.63	0.72
Fifth	0.52	0.42	0.49	0.57	0.47

Table 6: Longevity Change in Months from Universal Subsidy by Income and Wealth

Income	Wealth Quintile				
Quintile	Bottom	Second	Third	Fourth	Top
Bottom	0.10	0.02	0.21	0.47	0.40
Second	0.89	0.73	0.60	0.59	0.60
Third	0.78	0.51	0.57	0.50	0.52
Fourth	0.49	0.51	0.64	0.76	0.67
Top	0.57	0.58	0.45	0.39	0.59

Table 7: Lifetime Expected Cost of Universal Subsidy by Income and Health

Income		Range of Health h			
Quintile	All	$(0, 0.25]$	$(0.25, 0.5]$	$(0.5, 0.75]$	$(0.75, 1]$
Bottom	\$1,512	\$5,550	\$241	\$572	\$1,576
Second	\$1,796	\$208	\$286	\$1,362	\$6,863
Third	\$6,834	\$364	\$4,205	\$9,025	\$7,375
Fourth	\$13,821	\$5610	\$13,868	\$12,512	\$17,738
Top	\$172,464	\$76,548	\$77,324	\$211,027	\$171,104

Table 8: Lifetime Expected Cost of Universal Subsidy by Income and Wealth

Income	Wealth Quintile				
Quintile	Bottom	Second	Third	Fourth	Top
Bottom	\$8	-\$4	-\$10	\$365	\$5,940
Second	\$664	\$513	\$248	\$560	\$7,418
Third	\$709	\$1,968	\$2,226	\$4,485	\$23,211
Fourth	\$5,144	\$5,451	\$8,652	\$14,589	\$32,571
Top	\$25,331	\$41,428	\$68,350	\$119,515	\$560,414

Table 9: Longevity Change in Months from Targeted Subsidy by Income and Health

Income		Range of Health h			
Quintile	All	$(0, 0.25]$	$(0.25, 0.5]$	$(0.5, 0.75]$	$(0.75, 1]$
Bottom	0.35	0.00	0.08	0.51	1.20
Second	0.50	0.04	0.23	0.93	0.54
Third	0.03	0.00	0.05	0.03	0.02
Fourth	0.00	0.00	0.00	0.00	0.00
Top	0.00	0.00	0.00	0.00	0.00

Table 10: Longevity Change in Months from Targeted Subsidy by Income and Wealth

Income	Wealth Quintile				
Quintile	Bottom	Second	Third	Fourth	Top
Bottom	0.16	0.14	0.37	0.55	0.45
Second	0.75	0.57	0.51	0.45	0.25
Third	0.07	0.04	0.01	0.02	0.01
Fourth	0.00	0.00	0.00	0.00	0.00
Top	0.00	0.00	0.00	0.00	0.00

Table 11: Lifetime Expected Cost of Targeted Subsidy by Income and Health

Income		Range of Health h			
Quintile	All	$(0, 0.25]$	$(0.25, 0.5]$	$(0.5, 0.75]$	$(0.75, 1]$
Bottom	\$354	\$7	\$95	\$341	\$1,782
Second	\$431	\$44	\$98	\$643	\$1,047
Third	\$45	-\$1	\$27	\$50	\$165
Fourth	\$0	\$0	\$0	\$0	\$0
Top	\$0	\$0	\$0	\$0	\$0

Table 12: Lifetime Expected Cost of Targeted Subsidy by Income and Wealth

Income	Wealth Quintile				
Quintile	Bottom	Second	Third	Fourth	Top
Bottom	\$138	\$75	\$252	\$312	\$842
Second	\$216	\$156	\$318	\$333	\$1,046
Third	\$24	\$32	\$26	\$47	\$89
Fourth	\$0	\$0	\$1	\$0	\$1
Top	\$0	\$0	\$0	\$0	\$0

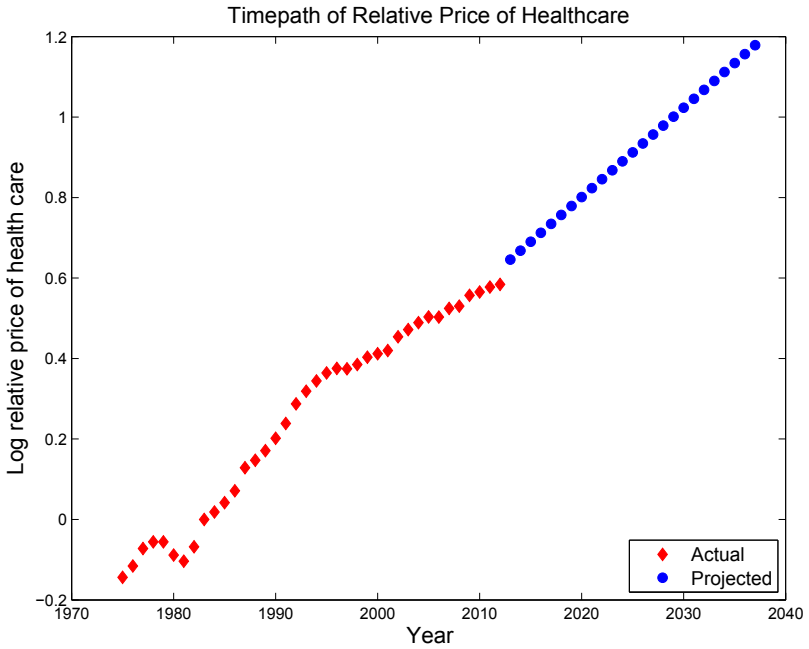

Figure 1: Log relative price of healthcare over time, actual and projected. Calculated as the medical component of the CPI divided by the non-medical component.

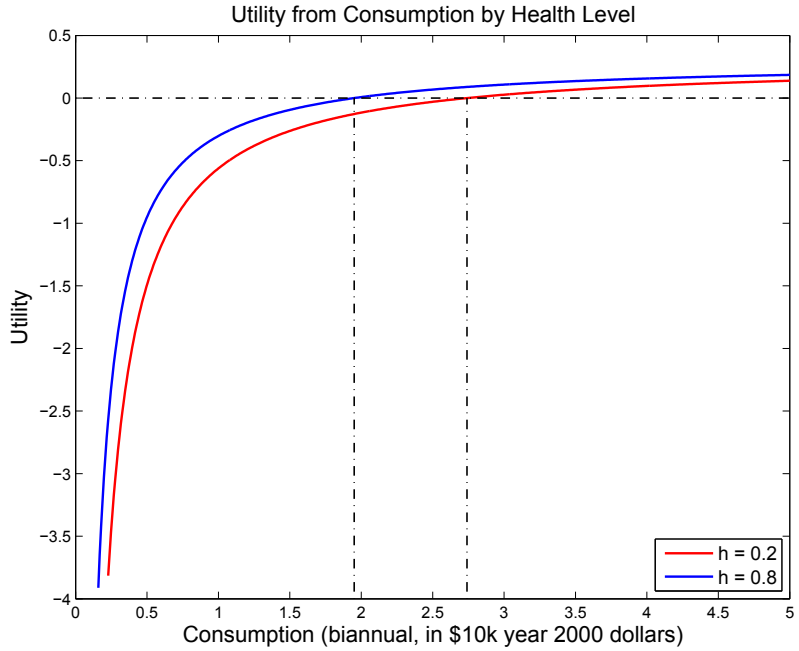

Figure 2: Utility function in good and bad health, assuming very low medical needs. Dashed lines indicate critical levels of consumption where $u(c, \cdot) = 0$.

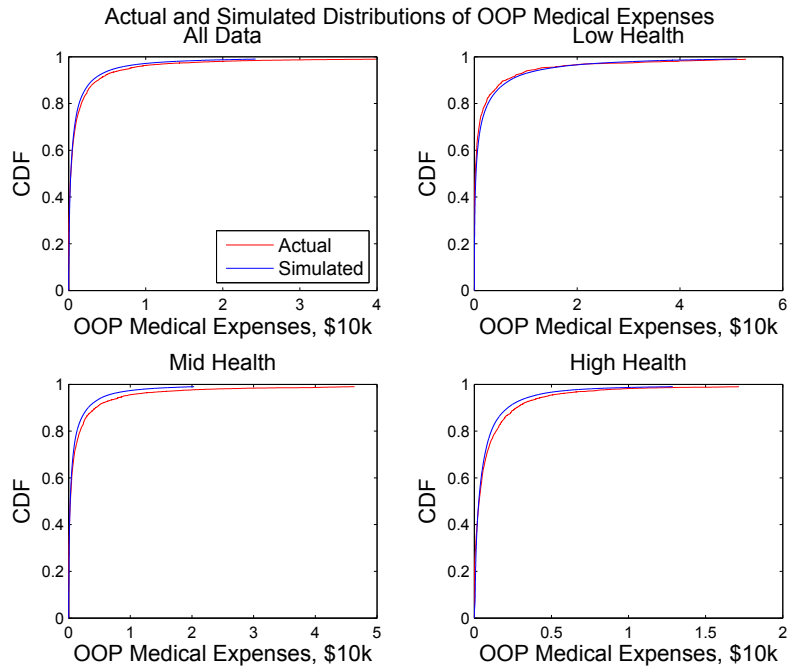

Figure 3: Actual and simulated distributions of out-of-pocket medical expenses for all individuals and for observations in low health ($0 < h \leq \frac{1}{3}$), medium health ($\frac{1}{3} < h \leq \frac{2}{3}$), and high health ($\frac{2}{3} < h \leq 1$).

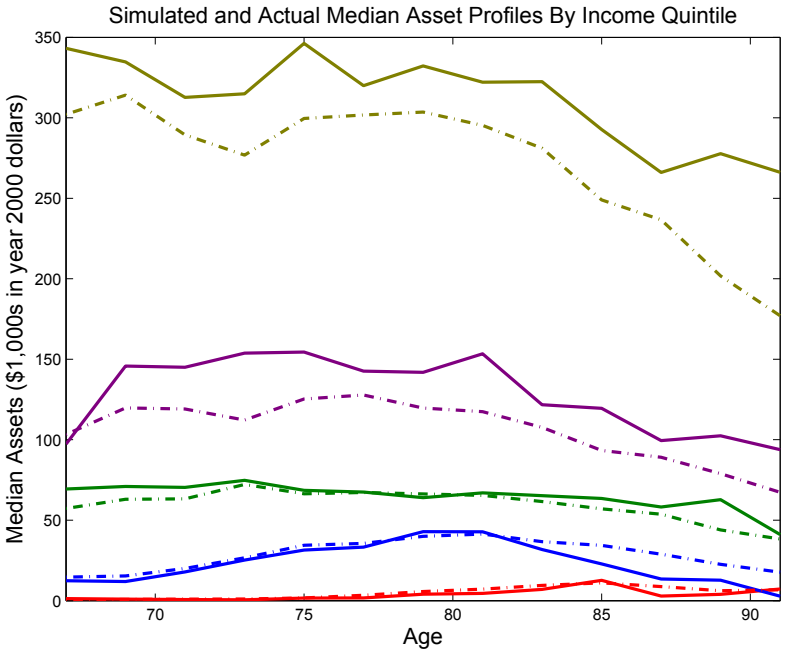

Figure 4: Actual (solid) and simulated (dashed) median asset profiles of individuals across income quintiles, age 67 to 91.

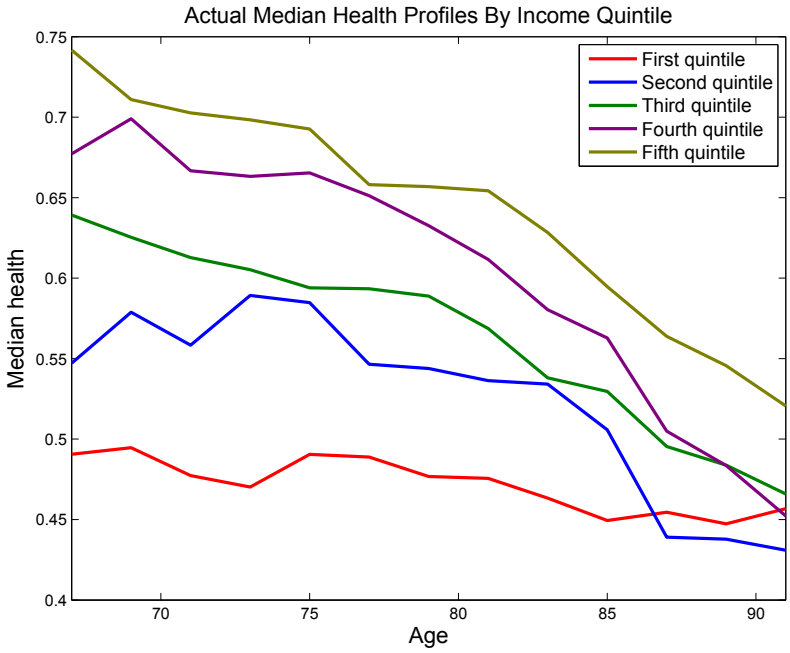

Figure 5: Actual median health profiles of individuals across income quintiles, age 67 to 91.

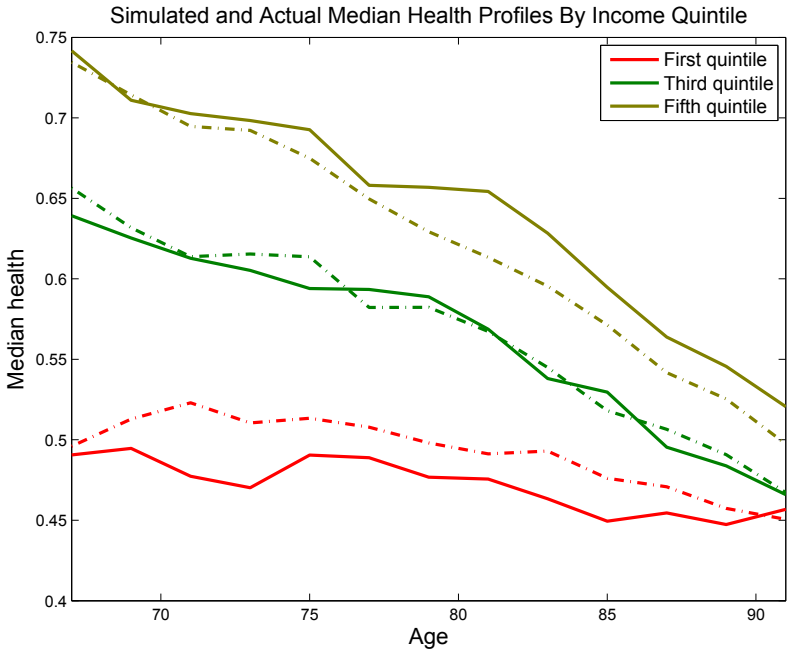

Figure 6: Actual (solid) and simulated (dashed) median health profiles of individuals across income quintiles, age 67 to 91.

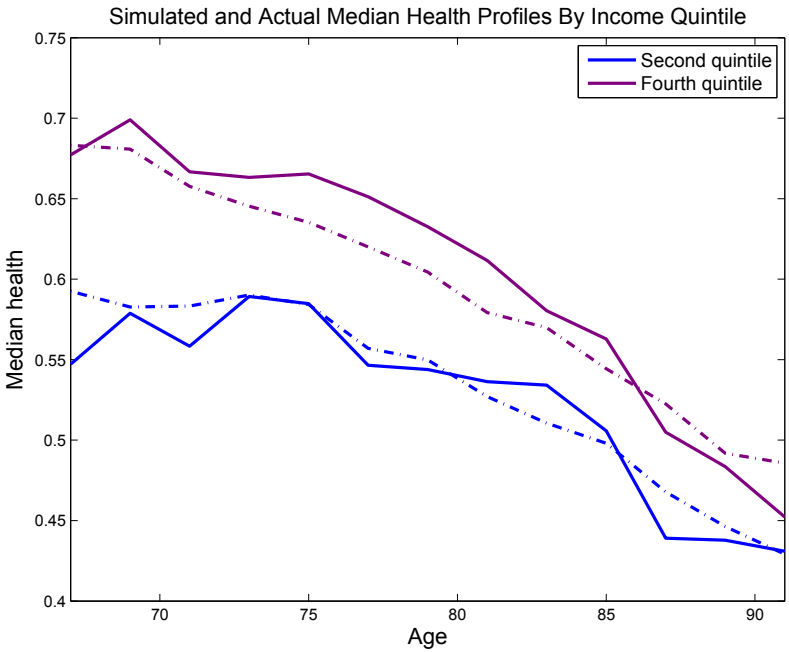

Figure 7: Actual (solid) and simulated (dashed) median health profiles of individuals across income quintiles, age 67 to 91.

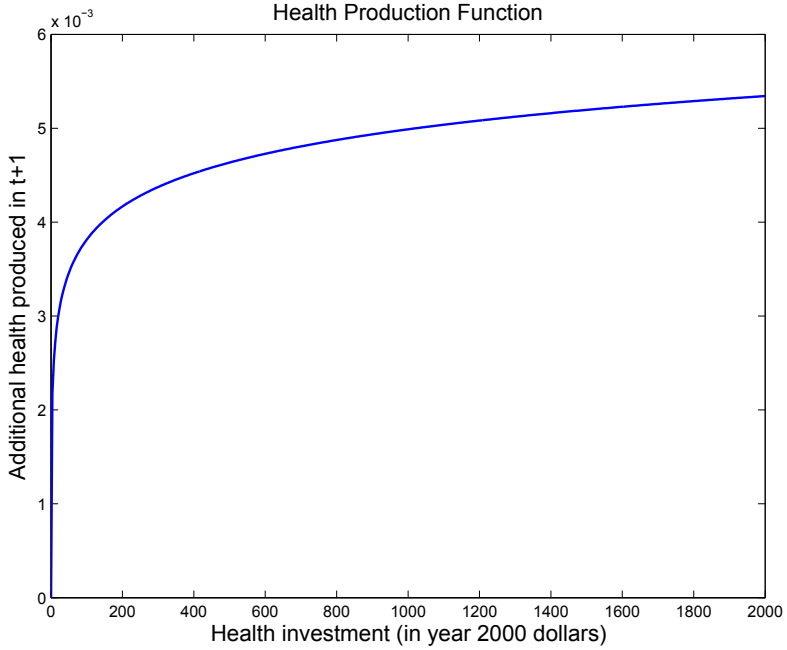

Figure 8: Additional health produced by dollar value of health investment purchased (at 2010 price level).

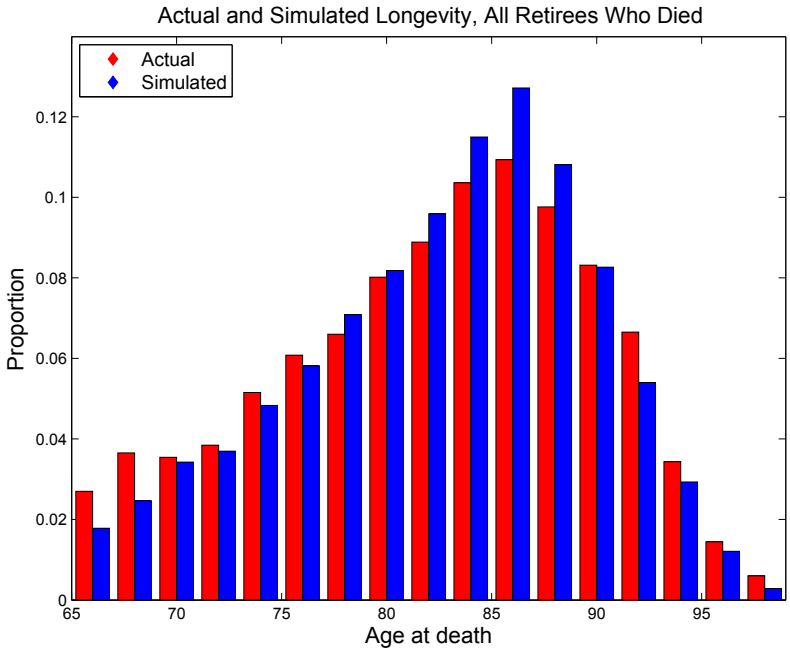

Figure 9: Actual and simulated distributions of longevity, in two-year blocks.

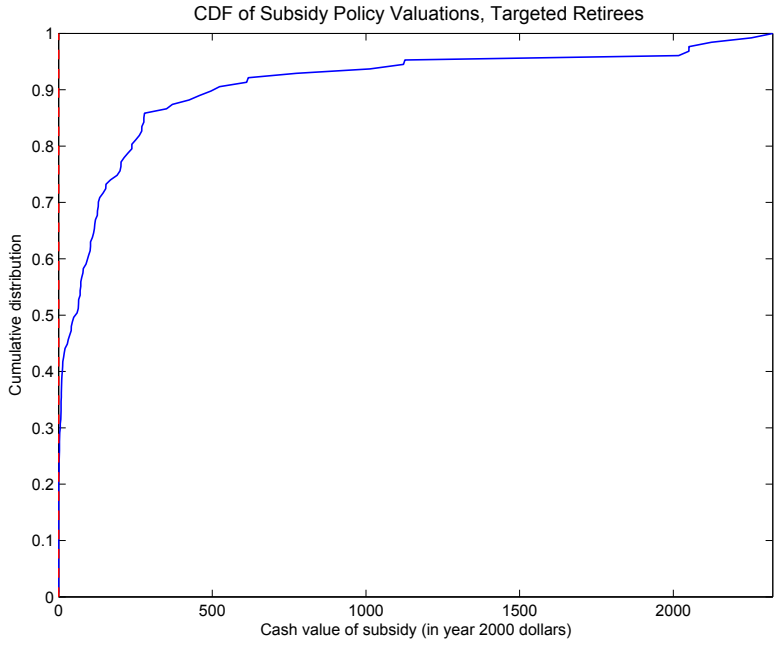

Figure 10: CDF of valuations of subsidy policy for individuals in first and second income quintile with $h > 0.5$ in 2010.

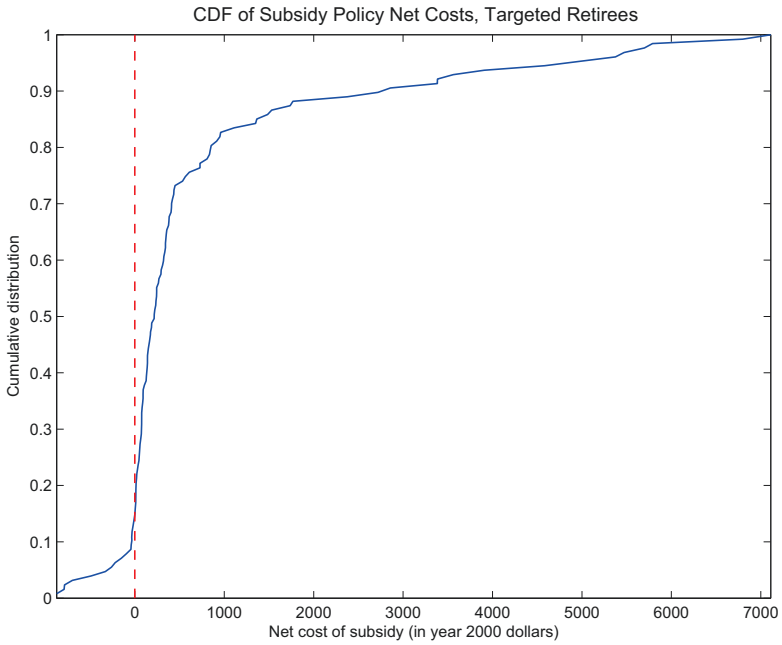

Figure 11: CDF of expected cost of subsidy policy for individuals in first and second income quintile with $h > 0.5$ in 2010.

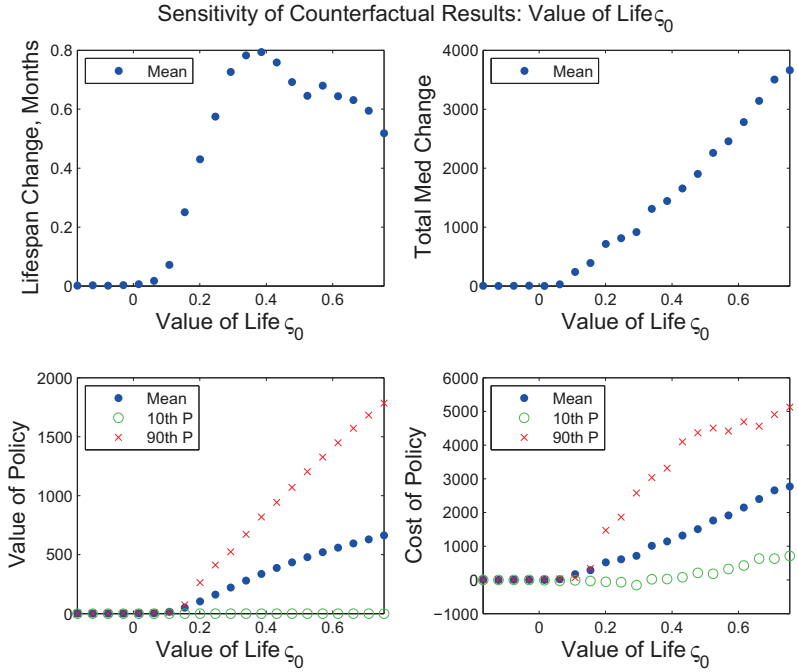

Figure 12: Sensitivity of counterfactual results to changes in the value of life, ς_0.

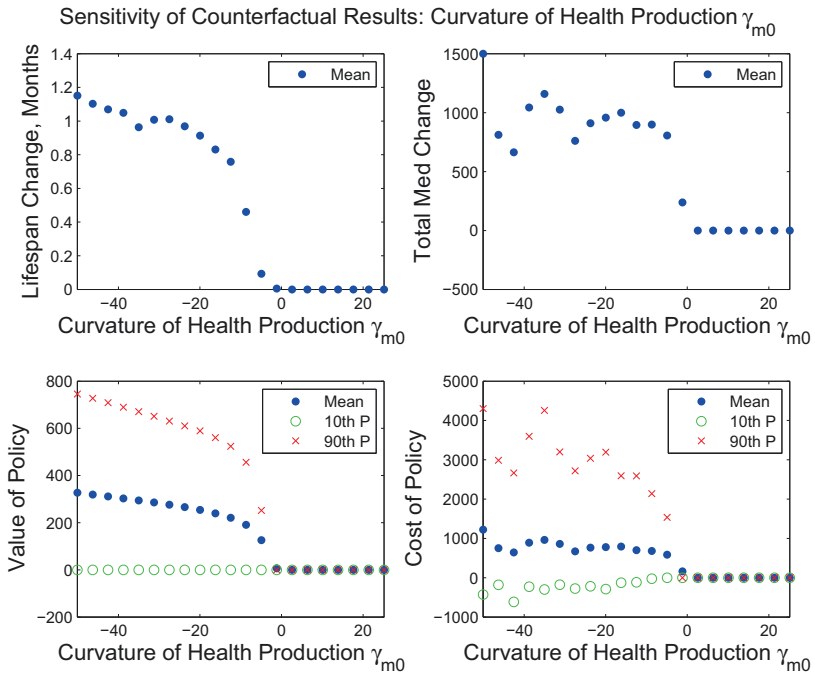

Figure 13: Sensitivity of counterfactual results to changes in the curvature of the health production function, γ_{m0}.

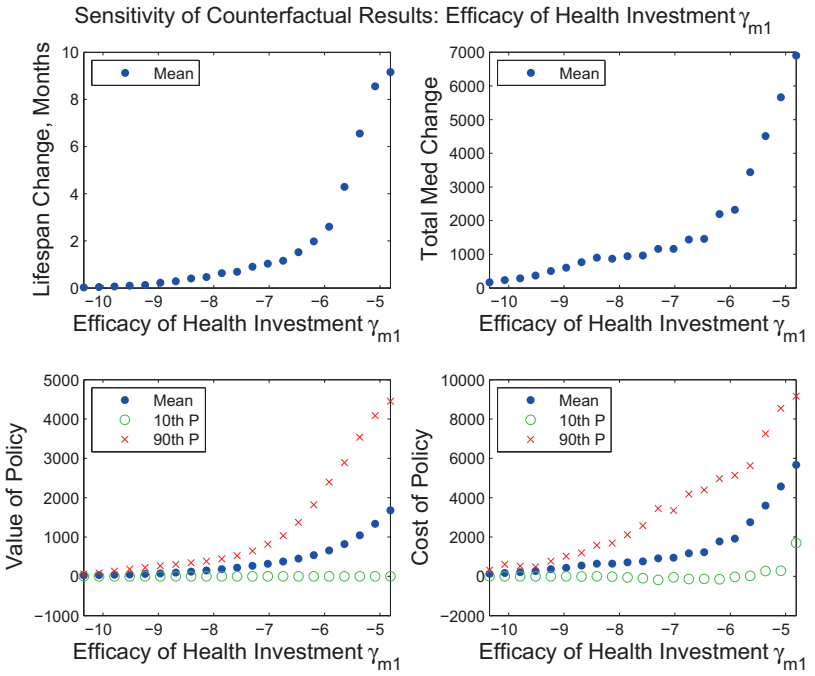

Figure 14: Sensitivity of counterfactual results to changes in the efficacy of health investments, γ_{m1}.

Appendices

A Income Profiles

To generate the 150 income profiles, individuals are first sorted into appropriate types by their sex, cohort (in two year blocks), and income quintile. Individuals' income quintile is determined by the rank of their average income over their first two observations relative to their sex-cohort peers. While this method is not perfect because individuals in the same cohort enter the data at different times, using more periods of data to sort by income quintile is potentially problematic due to survivor bias. The average growth rate of income at each age is calculated[6]; these values are used to calculate the average cumulative growth of income since ages 65-66. Each observation of income is normalized by the cumulative growth rate for that age. These "de-aged" income values are then averaged across observations within type to calculate a measure of permanent income for each type.

The income profiles are then constructed by the following method. First, for combinations of type and age with at least ten observations, the median of the relevant data is used as the the income profile value. Second, for age-type combinations that are older than the maximum (or younger than the minimum) filled by the first step, the income profile value is assigned by extrapolating the last age filled (or first age filled, for earlier ages) by the first step based on the average growth rates calculated above. Third, for age-type combinations that were not filled by the first step but are between the minimum and maximum ages for that type, the income profile value is assigned as the midpoint of the adjacent income profile values (all gaps are exactly one period). Fourth, for the handful of types that have no ages with at least ten observations (older male cohorts), the cumulative growth rates and measure of permanent income are used to construct the entire income profile.

B Simulation

Each qualifying individual in the HRS data is assigned to one of the one hundred fifty types based on their sex, cohort, and income quintile. The behavior of fifty copies of each individual are simulated from their entry into the dataset through 2010, the last period observed. Individuals' health and assets at the time of their entry into my sample are used as the initial conditions for the simulation. In each simulated period, the sequence of events is as described in Section 2.1: the individual receives a draw of his conditional distribution of medical needs, receives the appropriate income and pays

[6]For ages above 97 with very little data, I assume a growth factor of 0.972, the average growth rate from age 89-95.

insurance premiums, then makes decisions about purchasing goods as dictated by the behavioral functions just solved for, then receives a mortality and health shock to determine his next health state. The sample is replicated so that the data generated by this process more closely reflect the true distribution rather than a finite sample. To ensure convergence of the estimation, the underlying shocks applied to each simulated individual are held constant across evaluations of parameter sets.

C Moments

The categories of observations are combined in six different ways to create various kinds of cells: 175 income-wealth-year cells, 210 sex-income-health-year cells, 210 sex-cohort-year cells, 325 income-wealth-age cells, 25 income-wealth cells, and 30 age-sex cells. The range of health is also divided into upper and lower halves ($h \gtrless 0.5$) and into thirds for particular moments. Cells that would contain fewer than forty individuals (at least 30 living) are not used in the estimation. Not all moments are calculated for each cell (see below); overall, there are 1454 moments for the 32 model parameters.

Median-based moments use the same form as in DFJ, measuring the proportion of actual observations that are greater than the simulated median. For example, consider cell ι, which contains (say) observations of individuals in the third income quintile and second wealth quintile in year $t = 2002$. Defining $\bar{b}_\iota$ as the simulated median, the median assets moment for this cell is:

$$x_{\iota,b} = \sum_{it \in \iota} (\mathbf{1}(b_{it} > \bar{b}_\iota) - 0.5)/||\iota||. \tag{19}$$

Moments for median health and out-of-pocket medical expenses are defined similarly. At the true parameters, each of these moments has an expectation of zero, as half of the actual data should lie above the simulated median and half below. Because these moments are a discontinuous function of the model parameters, a linear interpolation is applied to correct for continuity (omitted above for clarity). The median-based moments are always on the interval $[-0.5, 0.5]$.

The non-median moments (mortality rate, the variance of out-of-pocket medical expenses, and the correlation of health status between periods) transform the ratio of the simulated and actual values onto the interval $[-0.5, 0.5]$. Calling the actual value from the data $\hat{d}_\iota$ and the simulated value $\tilde{d}_\iota$ for a particular cell, the mortality moment is calculated as:

$$x_{\iota,\theta} = \frac{1}{1 + \tilde{d}_\iota / \hat{d}_\iota} - 0.5. \tag{20}$$

The fraction in this equation ranges between 0 and 1 as the ratio of the simulated to actual value moves between infinity and zero; it equals one half when the values are identical, so the moment is zero when the simulation matches the data exactly. The mortality rate is calculated as the fraction of individuals eligible to be in a cell divided who have already died divided by the total number eligible to be in that cell; the variance of out-of-pocket expenses and the correlation of consecutive health states are calculated as expected.

Only some of the moments are calculated for each cell, as necessary for identification of the parameters. The median assets moments are calculated for the income-wealth-year cells and the sex-income-health-year cells. The median out-of-pocket expense moments are calculated for the income-wealth cells, age-sex cells, and by health thirds. The median health moments are calculated for the sex-income-health-year cells, the sex-cohort-year cells, and the income-wealth-age cells. The mortality moments are calculated for the sex-income-health-year cells and the cohort-sex-year cells. The variance of medical expense moments are calculated by health thirds. The health correlation moments are calculated for health halves. Discussion of the selection of these moments and how they identify the parameters can be found in Section 4.2.

D Weighting Matrix

Following standard methods, the second stage weighting matrix W is calculated based on the moments of the first stage estimate. The moments are divided into five sets based on their type: median assets moments, median medical spending moments, median health moments, mortality moments, and other moments. Using median assets as an example, the weighted variance for each of the moment types is calculated as:

$$w_b = \sum_{\iota \in I_b} (||\iota|| x_{\iota,b}^2) / \sum_{\iota \in I_b} ||\iota||, \qquad I_b = \{\iota | x_{\iota,b} \quad \text{exists}\} \tag{21}$$

The x values used above are the moment values calculated at the first stage estimate Δ_1. The weighted variance of all moments is used for the "other" moments, as there are only five moments that do not fit the other categories (three variance of medical spending and two health correlation moments). The second stage W is then constructed as a diagonal matrix with each element equal to the number of observations in that moment divided by the weighted variance of that moment type. The functional form of the non-median moments tends to yield a much smaller range of moment

values than the median-based moments, thus W puts more emphasis on these moments so that they are not dominated by the median moments.

E Standard Errors

The standard errors reported in Table 1 are calculated using the Jacobian of the moment function $X(\Delta)$ evaluated at the second stage estimate Δ_2. The jth column of the Jacobian (the derivatives of the moments with respect to the jth parameter) is numerically approximated as:

$$\nabla_j X(\Delta_2) = \frac{X(\Delta_2 + h_j) - X(\Delta_2 - h_j)}{2 \cdot 10^{-4} \Delta_2^j}, \qquad h_j = 10^{-4} e_j \Delta_2^j \tag{22}$$

In the above equation, e_j is the vector of zeros with a one in the jth index, and Δ_2^j is the estimate of the jth parameter. The covariance matrix of the parameter estimates is then calculated as:

$$\Sigma = ((\nabla X(\Delta_2))' \nabla X(\Delta_2) W)/T \tag{23}$$

T is the total number of observations of living individuals in the data. The standard errors reported are the square roots of the diagonal elements of Σ.

This page intentionally left blank.

Chapter 2:
The Role of Medical Inflation in Asset Decumulation and Demand for Medical Care

Abstract

This chapter uses the estimated model from the previous chapter to investigate how medical inflation affects the consumption, medical care, and saving decisions of retired Americans. I simulate the behavioral decisions of a sample of young retirees (65-70) in the Health and Retirement Study from 2010 until their deaths under the average rate of medical inflation over the past three decades and alternate high and low rates. Further, the effects of medical inflation are decomposed into two parts: the intratemporal or reactionary effect, motivated by experiencing changes in the current price of care; and the intertemporal or precautionary effect, arising from foreseeing changes in future price growth. I separate these effects by simulating an environment where medical inflation has changed but individuals do not adapt their beliefs, and an environment where individuals believe inflation has changed but it continues normally. The simulations reveal that high medical inflation is associated with retirees retaining more wealth into old age as a buffer stock against medical costs with greater variance, rather than wealth being more rapidly depleted by the larger costs. The increased asset retention is made possible by cuts to consumption, which are relatively steady for the remainder of individuals' lives. The decomposition shows that changes to consumption represent a balanced tension between opposing precautionary and reactionary effects, while the effect on medical care is driven almost entirely by the reactionary effect. More strongly, medical spending behavior is determined mostly by the present price of care itself, rather than its signal of future prices.

1 Introduction

For the past three decades, the United States has experienced considerable medical inflation– the prices of health care goods and services growing at a faster rate than other consumer goods. From 1981 to 2011, the medical component of the consumer price index (CPI) had an average growth rate 2.24% higher than the non-medical component. While medical inflation has occurred with near perfect consistency (with positive values in all but two years in the thirty year span), there has been considerable variation in the medical inflation rate from year to year. As seen in Figure 1, after thirteen years of very high medical inflation averaging 3.5% and no single year significantly below 2%, the relative price growth of medical care moderated considerably from 1994 to 2000, averaging barely over 1% and no single year over 2%. In the past decade the average rate has returned to levels closer to (but still slightly below) the historic average, with medical inflation particularly low in the past four years. It has been suggested that the most recent slowdown is largely due to the Great Recession and other transient effects, and that a re-acceleration is on the horizon, but this is in dispute. Given its history, there is considerable uncertainty about the future path of the relative price of medical care. Further, the Congressional Budget Office projects that federal spending on health care, primarily through Medicare and Medicaid, will be the biggest driver of future U.S. budget deficits. While the ongoing growth in health care spending has several sources (including an aging population that requires more medical services and a wealthier population that prefers more health care to increase longevity), medical inflation is widely considered to be a key factor.

This chapter examines how medical inflation affects the decisions of retired Americans concerning consumption, medical care, and saving. In a dynamic environment in which individuals foresee changes in the conditions they will face in the future, there are effects that could push each of these objects in different directions. For example, the greater variance of future medical expenses will motivate individuals to retain more precautionary savings, while actually experiencing the larger costs will deplete their savings more quickly. From a purely theoretical perspective, it cannot be known which of these two effects dominates, nor the timing of a switch in dominance. To this end, this chapter applies the structural model estimated in the previous chapter to simulate the effects of changes in medical care on the behavioral choices of the elderly. Using a sample of single, retired Americans between the ages of 65 and 70 years in the 2010 wave of the Health and Retirement Study (HRS), I solve the retired lifecycle problem for a variety of types of individuals under different rates of medical inflation and individuals beliefs about this rate. To better understand the motivation

for the behavioral changes caused by medical inflation, the analysis focuses on decomposing the effects into reactionary (or intratemporal) and precautionary (or intertemporal) portions. That is, the effects that arise due to changes in the price level from ongoing inflation versus effects from anticipated price growth, respectively.

The structural model estimated in the previous chapter is particularly suitable for conducting this analysis. Most importantly, it includes most of the major components of retirees' economic behavior, including consumption, saving, high variance medical expense shocks, and endogenous health evolution; however, it omits the composition of assets, treating all forms of wealth as a unitary asset. The parameters of the model were estimated to generate simulated life cycle trajectories that match the data for these objects not just for the population as a whole, but disaggregated by health, income level, relative wealth, and cohort. As shown in the last chapter, the model fit was very good considering that the estimation was extremely overidentified. While it was not the focus of the estimation, the model was implicitly calibrated to account for medical inflation by separately matching the medical expenses of individual birth cohorts, each of which experiences a different timepath of medical prices as they age. By taking a structural approach and uncovering so called "deep parameters" that govern the preferences and underlying processes several steps removed from observable data, rather than employing a reduced form descriptive method, the model can be used to counterfactually simulate the effects of changes in medical inflation and produce credible results. More strongly, structural modeling is uniquely suited to disentangling the motivation behind these effects, separating the overall effect into reactionary and precautionary components.

1.1 Discussion and Results

To elucidate this decomposition, consider again the example of retirees' saving decision. The fact that the growth rate of the price of medical care has increased will cause individuals to hold more assets as a precautionary buffer against large medical expense shocks. Conversely, the higher price levels as they are actually experienced will put downward pressure on savings. Further, there may be intertemporal substitution effects due to higher medical inflation. In the classic Grossman model of medical care as an investment in a depreciating capital stock of health, faster price growth of medical care causes larger health investments in the present, as it will be more expensive to replenish health in the future. As time passes and the higher prices actually manifest, however, health investment should fall. Moreover, the marginal value of future wealth is higher (while its absolute level is lower)

because of larger medical expense shocks and potentially lower levels of consumption, decreasing the returns to health investment. A similar conflict arises for consumption, discussed fully in Section 3.2.

The simulated decomposition seeks to separate these effects and differentiate between changes that occur because of the *price growth rate* for medical care versus changes due the *price levels* that flow from it. To this end, the main analysis focuses on three counterfactual simulations. First, the complete effect of changes in medical inflation, in which medical prices grow at a more rapid pace and individuals become aware of this change just before it happens. Second, the reactionary or intratemporal effect, in which medical inflation is faster than before, but individuals believe it has remained at the historic baseline rate; they correctly see the price of care as it occurs each period, but are always surprised by it. Third, the precautionary or intertemporal effect, in which individuals believe that inflation is higher than the baseline rate, but it is actually unchanged; they prepare for a worse future that never comes, and are also surprised by the price each period.[1] The latter two simulations represent a nearly perfect decomposition of the total effect of higher medical inflation: when their effects (on consumption, savings, medical care, etc) are represented as percentage deviations from the baseline quantities under the historic average rate of inflation, they sum to the total effect with almost no error. This decomposition is valid both in the short run and dynamically until very late ages in life, with slight differences only apparent as the few survivors approach 100 years of age.

The reactionary effect is not purely intratemporal, depending only on the change in the current price of medical care with no perceived change in future prices. Even for an individual who believes that medical inflation has not increased, the fact that the price today is higher than expected signals that future prices will also be higher than previously expected. That is, the individual believes that prices will continue to grow at the baseline rate, but now from a higher initial level. The reactionary effect thus comprises the pure intratemporal price effect as well as the signal of the current price on future prices. To separate these effects, I perform static decompositions of the own price elasticity of medical care and the cross price elasticity of consumption. Focusing only on changes in quantity demanded in the current period (2010, the most recent data wave), I solve the life cycle problem under three different paths of future prices for each of several alternate initial price levels. First, the total change, in which the initial price is altered from the true level and individuals believe medical

[1] Here, "surprise" means that the current price of medical care does not align with what the individual had believed it would be. Individuals in the model account for the new information and form new expectations about the future stream of medical prices, based on their belief about medical inflation.

inflation will continue from this alternate level. Second, the present change, in which the current price is altered, but individuals treat it as a temporary shock– that medical prices will return to their previous growth path. Third, the future change, in which the current price is unaltered, but individuals foresee that medical prices will jump to the alternate level next period and grow steadily from that level. Just as with the dynamic simulation, this division of single period behavior forms a nearly perfect decomposition of the static demand curves, separating the reactionary effect into the pure intratemporal price effect from the signal effect.

The main counterfactual simulation reveals that when retired individuals learn that medical inflation will be permanently elevated to the historically high rates seen in the 1980's and early 90's (about 3.5%), they immediately cut purchases of all goods (consumption and both medical care goods) by 0.4% to 0.8% relative to their behavior at the baseline inflation rate (about 2.25%). Foreseeing larger future medical expenses, individuals begin to retain more wealth as a buffer against these shocks. As time passes and the higher prices of medical care actually arrive, spending on medical care rapidly increases relative to baseline even as quantities are cut (but at a slower rate). Despite the large increases in medical expenses, retirees continue to save more than in the baseline scenario due to consumption remaining persistently half a percent lower. Even when increased medical expenses overtake decreased consumption to boost total spending above baseline levels, wealth continues to climb relative to baseline through higher investment income. That is, the desire to hold additional precautionary wealth overcomes the downward pressure from higher prices. The simulations also reveal that demand for health investment is significantly more elastic than for medical consumption (representing health care that mitigates medical conditions, rather than curing or preventing them). This seems to be attributable to health investment being a superior good in the model; as medical inflation erodes the purchasing power of retirees' income, health investment is particularly sensitive to these changes. Moreover, the depressed levels of consumption make additional periods of life less attractive, reducing the returns to health investment. An alternative simulation in which individuals face historically low medical inflation (about 1.1%, as in the mid 1990's) shows a near perfect mirror image of these effects, indicating an aggregate linearity to a model that is highly non-linear on an individual level.

The decomposition of the effects of medical inflation into reactionary and precautionary effects have significantly different patterns across the variables of interest. The relative accumulation of wealth is initially spurred only by retirees' precautionary motive as they foresee higher prices, but is eventually supported by the reactionary effect of actually experiencing the higher prices. As

higher medical prices themselves will tend to deplete wealth, the reactionary effect on wealth is thus dominated by the signal of even higher prices in the future, rather than present medical costs; individuals will save more when faced with higher medical inflation *even if* they don't recognize that it is happening. The simulations also show that the steadily lower level of consumption reflects a balance between the precautionary and reactionary effects. While the former induces an immediate cut followed by a gradual increase relative to baseline levels (as assets accumulate and finance higher consumption), the latter causes a steady decline in consumption as individuals find themselves repeatedly surprised and unprepared for higher prices each period. These effects nearly perfectly balance out to a steady level relative to baseline until very late ages. In contrast, purchases of both medical care goods are driven almost entirely by the reactionary effect of higher prices; the precautionary effect generates a similar pattern of preparatory cuts followed by gradual increase, but the magnitude of the changes is very small in comparison. Moreover, the static decomposition of the own price elasticity of medical care shows that the reactionary effect is almost entirely attributable to the pure intratemporal price level effect rather than the current price's signal of future prices. Conversely, the cross price elasticity of consumption is dominated by the signal value, reinforcing the conclusion that changes to consumption are the primary mechanism by which individuals adjust their savings in response to changes in medical inflation. Finally, the prediction of the Grossman model that faster medical inflation will induce greater health investment through an intertemporal substitution effect does not seem to hold in a model with endogenous savings; the effect may exist, but it is overwhelmed by precautionary concerns.

The findings of this chapter hold relevance for researchers who seek to structurally model and estimate a dynamic general equilibrium framework that endogenizes the price of medical care based on aggregated choices of individuals and households. The ideal for any dynamic model is for all agents to have rational expectations about the future– that there is a consistency between agents' expectation of future states of the world, their actions in response to those beliefs, and the future states generated by those actions both individually and collectively. However, meeting this ideal requires repeated iterations (often nested) of generating beliefs, solving for individually optimal behavior, aggregating behavior into macroeconomic outcomes, and generating new beliefs based on these outcomes until consistency is reached. In non-linear structural models that also include heterogeneity among agents, the computational burden of estimating the model can become excessive. For models concerned with the price of medical care, this chapter provides evidence that it is reasonably safe to employ shortcuts rather than attempting a full rational expectations framework. That

is, a solution method that assumes a particular macroeconomic future will not produce a drastically different result for individuals' optimization problem than would a rational expectation model. Because demand for care depends almost entirely on its current price and demand for consumption is fairly insensitive to the price of care, reasonable assumptions about future medical prices will provide a sufficiently accurate approximation to a general equilibrium model that endogenizes medical inflation.

This chapter focuses on the effects of medical inflation on the behavior of only single retired individuals, as the structural model was estimated on this population in order to simplify the solution to the life cycle problem. However, the conclusions of the counterfactual simulations are likely broadly applicable more generally, including to working individuals and households with more than one member. None of the main results seem to be dependent on the assumptions inherent to the sample restriction to single, retired individuals, but there would be some differences. Medical care represents a much smaller portion of total spending for younger individuals, and thus overall spending growth would likely be slower for the working population. This would allow savings to accumulate even more rapidly, a necessary reaction given that the households expect to survive for more periods and thus face even higher medical price levels than the individuals in the simulations. The increased saving rate would undoubtedly be financed by cuts to consumption, but there could be greater heterogeneity of these effects across income and wealth than seen in the retired population (see Section 3.4), as asset accumulation rates exhibit extreme variation by household income. In a macroeconomic model founded on these individual behaviors, the expectation of higher medical inflation could lead to depressed aggregate demand while boosting the supply of financing for capital investment, potentially affecting economic growth. Until this model is extended and incorporated into a larger general equilibrium framework, however, these effects are only speculative.

1.2 Related Literature

As mentioned above, persistent growth in medical costs relative to the entire economy can be attributed to several sources: changes in the demographic composition of the United States as the population ages, growth in real income that shifts the composition of goods and services demanded, changes in population insurance status, technological developments in medicine over time, and price growth of individual medical goods and services. In a history of medical spending growth from 1960 to 1993, Peden and Freeland (1995) find that only 30% of the growth in this time frame was due to

changes in demographics, income, and insurance; the remaining 70% is attributed to technological development and pure price growth. Levit et al. (2002) find that changes in insurance contracts played a larger role in the late 1990s, as a tight labor market and a backlash against managed care led to more generous plans and subsequent increase in demand. Moreover, spending growth was particularly driven by pharmaceuticals (through new brand-name drugs) and hospital care (by dint of its large share of total care). Examining Medicare cost growth specifically, Thorpe and Howard (2006) find that the sickest patients– those who are being treated for five or more conditions– account for nearly all growth between 1987 and 2002. In related work, Thorpe, Florence, and Joski (2004) show that a handful of the most costly medical conditions drive about half of total medical spending growth, and that this growth is mostly from a rise in the rate at which the population is treated rather than increases in treatment cost. Bundorf, Royalty, and Baker (2009) consider growth among privately insured individuals from 2001 to 2006 and find that it is mostly attributable to changes in quantity, with price growth more of a factor in drug costs. Looking at the same population over a similar time frame, a very recent working paper by Dunn, Liebman, and Shapiro (2014) decomposes medical spending growth into four components: service price, treated prevalence, service utilization (intensity), and demographics. They find that price growth accounts for half of total medical spending growth, but only 10% when the prices are deflated by a consumer price index for all other goods and services.

Price indexes are usually measured by holding fixed a representative market basket of goods and services and tracking their total cost over time. In the case of medical care, however, the mix of services purchased changes much more rapidly than in other areas due to rapid technological development. If newer services are less expensive than older ones and consumers substitute into these services as they become available, then traditional price indexes might overstate medical inflation by not accounting for these changes. Indeed, Aizcorbe and Nestoriak (2011) find that medical inflation is reduced by nearly half when a so-called medical care expenditure index (MCE), tracking the cost of treating a specific disease, is used rather than a traditional service price index (SPI). For example, an MCE considers a good to be "treatment for a typical episode of depression" rather than "a one hour talk therapy session" or "a one month supply of an SSRI". Using a much larger dataset, Dunn, Liebman, Pack, and Shapiro (2012) similarly find that MCEs exhibit a lower rate of inflation than SPIs, although not as extreme a difference: about one fourth or one fifth slower. A working paper by the same authors, Dunn, Liebman, and Shapiro (2013) finds that the different inflation rates as measured by an MCE vs SPI approach are much more similar when the service prices are calculated

on a by-procedure basis rather than by-encounter.

While the inflation rates used in this paper are based on the medical component of the CPI, computed by the Bureau of Labor and Statistics using a SPI methodology, this price index might not best fit what the model is trying describe. As will be seen in section 2.1, individuals in the model purchase medical care to reduce the utility loss from their current period medical needs, drawn from a continuous distribution. That distribution represents an approximation to some time-invariant set of discrete conditions and illnesses that individuals may want to treat. The quantity of medical care they purchase represents whatever mix of services are used to treat that condition at that time, whether or not those services are the same as in earlier periods. The model thus seems to care about MCE-based prices– how much will it cost to reduce the utility loss for a particular illness at this time– rather than the SPI-based prices that were used to calibrate the model. As there is still debate over whether the two approaches yield substantially different rates of medical inflation, this might not adversely affect the validity of the simulations. To the extent that medical inflation has been mismeasured and/or misapplied, the qualitative results of the simulations are still of interest even if the magnitudes should be scaled down.

In section 3.6, I perform a static decomposition of the price elasticity of demand for care to separate the pure intratemporal effect from forward-looking effects based on the signal value of the current price. This exercise shows that individuals in the model exhibit an elasticity of about $\epsilon = -0.462$, most of which ($\epsilon = -0.434$) is due solely to the current price. This elasticity is on the higher end of the range of previous estimates (in absolute value), but not unreasonably so. A 2002 study by the RAND Corporation compiled sixteen earlier studies of the price elasticity of care, with an incredible range of -0.02 to -2, though the bulk of the estimates were between about -0.1 and -0.3. The measures of both quantity (intensive vs extensive margin) and price (coinsurance rates vs copayments vs full service price) varied across the studies, as did the medical service studied (hospital stays vs prescription drugs vs mental health care); methodology also ran the full gamut from observational data collection to natural experiments to experiments (the famous RAND HIE). This chapter might be the first study to calculate elasticity through a structurally estimated model, but still finds a reasonable value that broadly agrees with earlier work.

2 Model

This section summarizes the model from the previous chapter and extends it to include a broader array of beliefs about future prices of medical care. Individuals, indexed by i, represent unmarried retired persons over the age of 65 living in the United States. They are lifetime expected utility maximizers over a finite but uncertain lifetime, with a common utility function and an intertemporal discount factor δ over discrete time t (in two-year periods). Conditional on survival to period t, the individual receives income y_{it} according to his exogenous characteristics. Beyond the individual's exogenous characteristics (or "type"), including sex, cohort, and income quintile, individual i's state entering time t is characterized by his net real assets b_{it} and his $h_{it} \in (0, 1]$ (a dead individual would have $h_{it} = 0$).

2.1 Timing, Prices, and Utility

Each period, the individual purchases non-negative quantities of three assets: composite consumption c_{it}, medical consumption μ_{it}, and health investment κ_{it}. The price of medical consumption and health investment is p_t, while the price of consumption is normalized to 1; the timepath of p_t is exogenous. All individuals correctly observe p_t at time t, but might have incorrect beliefs about future prices of medical care. To distinguish between the actual prices that occur and individuals' beliefs about those prices, p_t will always represent the true price at t, while $\hat{p}_{ts}$ will stand for individuals' belief at time t about the price of medical care in future period s ($s > t$). All individuals have common beliefs about the constant rate of medical inflation[2] that will arise in the future, $\hat{\pi}$, which might differ from the true rate of inflation π. Starting from an initial time $t = 0$, the path of actual medical care prices is:

$$p_t = \pi^t p_0, \qquad t \geq 0. \tag{1}$$

Individuals' belief at time t about the price of care at future time s is:

$$\hat{p}_{ts} = \hat{\pi}^{s-t} p_t, \qquad s \geq t. \tag{2}$$

Each period, the individual first receives income y_{it} and pays medical insurance premiums z_{it}, then his medical needs shock η_{it} is realized. Given the shock, he chooses optimal quantities of the

[2] "Medical inflation" here refers to the rate of increase in a composite index of medical care prices greater than the rate of increase for all other consumer goods. This is convenient in the current context because the price of composite consumption is normalized to 1. Other sources may refer to this same concept as "excess medical inflation" or similar terms.

three goods subject to his budget constraint. Finally, his health state evolves and the next period begins. Current period utility flow is given by:

$$u(c_{it}, \mu_{it}; \eta_{it}, h_{it}, h_{it-1}, b_{it}) = \begin{cases} (1+\alpha_1 h_{it})\frac{c_{it}^{1-\rho}}{1-\rho} + \eta_{it}\frac{\mu_{it}^{1-\nu}}{1-\nu} + \varsigma_0, & \text{if } h_{it} \in (0,1] \\ \omega_1 \frac{(b_{it}+\omega_0)^{1-\rho}}{1-\rho}, & \text{if } h_{it} = 0 \text{ and } h_{it-1} > 0 \\ 0, & \text{if } h_{it} = 0 \text{ and } h_{it-1} = 0. \end{cases} \tag{3}$$

The utility function for a living individual is given by the first case, the sum of two constant relative risk aversion terms. The relative marginal utility weights of consumption and medical consumption are determined by the current health state and the medical needs shock. When the coefficient of relative risk aversion for medical consumption ν is greater than one, utility from medical consumption is always negative– a penalty to utility from current medical needs that is tempered with medical care. The magnitude of η_{it} determines the amount of medical care that the individual will want to purchase to mitigate his current medical needs. The parameter ς_0 is a level shifter that determines the relative value of being alive; without it, flow utility (and lifetime expected value) would be strictly negative (worse than death), obviating the availability of health investment for nearly all individuals.

The second and third cases of (3) respectively represent the bequest motive and the normalization of the utility of death. In the period immediate following death (i.e. when previous health was positive but now is zero), the newly deceased receives a "scrap utility" payment depending on his total assets, with parameters ω_0 and ω_1 adjusting the curvature and scale of the scrap utility relative to ordinary consumption utility. As is typical, the bequest motive serves to prevent individuals from rapidly consuming their resources when death is likely imminent, allowing the model to better match observed behavior. The normalization of the utility of being dead to zero is necessary in a model where the timing of death is affected by the choice of health investment. Unlike many other dynamic models, affine transformations of the utility function are not benign when the number of periods is endogenous.

2.2 Medical Needs and Health Transitions

The medical needs shock η_{it} is drawn from a Weibull distribution whose scale parameter depends on the individual's sex, age, and health and shape parameter is a linear function of health. Formally:

$$\eta_{it} \sim f(\eta|sex_i, age_{it}, h_{it}) = \frac{k_{it}}{\lambda_{it}} \left(\frac{\eta}{\lambda_{it}}\right)^{k_{it}-1} e^{-(\eta/\lambda_{it})^{k_{it}}}, \qquad k_{it} = \beta_{k0} + \beta_{k1} h_{it}, \tag{4}$$

$$\lambda_{it} = \exp\left(\beta_0 + \beta_s sex_i + \beta_{a1} age_{it} + \beta_{a2} age_{it}^2 + \beta_{h1}(1-h_{it}) + \beta_{h2}(1-h_{it})^2\right). \tag{5}$$

This form allows for medical needs to become larger on average and have a greater variance as health declines without restricting the relationship between the mean and variance. Characteristics exogenous at time t will be summarized by the vector $Z_{it} = (sex_i, age_{it}, h_{it})$.

The individual's health evolves from period to period according to a stochastic process with two shocks. First, the individual receives a mortality shock, where survival probability depends on age, sex, and health. Second, individuals who survive the mortality shock then receive a health shock (which could also result in death with a sufficiently bad draw). The mean of the health shock depends on current age, sex, health, and health investment, while the variance depends linearly on health. The joint health evolution process is given by:

$$h_{it+1} = \begin{cases} \max\{0, \min\{1, \hat{h}_{it+1}\}\}, & \text{if } \Theta_{it} \leq 0 \text{ and } h_{it} > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

$$\Theta_{it} = \theta_0 + \theta_s sex_i + \theta_{a1} age_{it} + \theta_{a2} age_{it}^2 + \theta_{h1}(1-h_{it}) + \theta_{h2}(1-h_{it})^2 + \epsilon_{it}^{\theta}. \tag{7}$$

$$\hat{h}_{it+1} = \gamma_0 + \gamma_s sex_i + \gamma_{a1} age_{it} + \gamma_{a2} {age_{it}}^2 + \gamma_{h1} h_{it} + \gamma_{h2} h_{it}^2 \tag{8}$$

$$+ \exp(\gamma_{m1} + \gamma_{m2}(1-h_{it}))(\log(\kappa_{it} + \exp(\gamma_{m0})) - \gamma_{m0}) + (\sigma_0 + \sigma_1(1-h_{it}))\epsilon_{it}^{h}.$$

Both ϵ_{it}^{h} and ϵ_{it}^{θ} are standard normal error terms. Equation (7) gives the probit for the mortality shock, while (8) shows the distribution of health for individuals who survive the mortality shock. The first line of (8) gives the expectation of next period's health based on age, sex, and current health, while the first term of the second line provides the contribution to future health from health investment κ_{it}.[3] It is assumed that individuals cannot live past age 105, so the health distribution beyond this age is degenerate with unit mass at $h_{it+1} = 0$. The distribution of future health

[3] For the sake of brevity, a complete description of the functional form has been omitted. See previous chapter for more detail.

generated by (6), (7), and (8) will be denoted by $g(h_{it+1}|\kappa_{it}, Z_{it})$.

2.3 Solving the Model

Each period, the individual is thus tasked with choosing how best to divide his resources among consumption, medical consumption, health investment, and saving. That is, he must balance the immediate utility of composite and medical consumption against the future benefits from health investment and the need to save assets as a buffer against future uncertainty in medical needs and health. He must do so according to a typical budget constraint:

$$b_{it+1} = R_t \left(b_{it} + y_{it} + w_{it} - z_{it} - c_{it} - q_{it} p_t (\mu_{it} + \kappa_{it}) \right) \geq 0. \tag{9}$$

That is, next period's assets equal this period's money resources (assets plus income and welfare) less insurance premiums and the cost of his purchased goods. The individual pays only a fraction q_{it} of the total cost of his health care, according to the copay of his insurance. All individuals are subject to a utility floor $\underline{u}$ set by policy: if it would be impossible for current flow utility to achieve at least $\underline{u}$, the individual receives a welfare payment $w_{it} \geq 0$ that gives him exactly enough money to have a utility of $\underline{u}$ while purchasing no health investment.

The individual's problem is solved by backward recursion. Assume that next period's value function $V_{it+1}(b, h)$ is known, representing the discounted sum of expected utility the individual will receive from period $t+1$ until death if he begins the period with assets b and health h and makes the optimal choices from that period onward. Once the period t medical needs shock η_{it} has been realized, the individual's problem is described by:

$$V_{it}^{\bullet}(b_{it}, h_{it}, \eta_{it}) = \max_{c_{it}, \mu_{it}, \kappa_{it}} \left[u(c_{it}, \mu_{it}; \cdot) + \delta \int_0^1 V_{it+1}(b_{it+1}, h) g(h|\kappa_{it}, Z_{it}) dh \right] \text{ s.t. (9).} \tag{10}$$

The bullet denotes that this is the "intermediate value function", or the expected lifetime payoff for i from receiving a medical needs shock of η_{it} after he enters period t with the particular level of assets and health. This can be integrated with respect to the shock distribution in order to find the value function at the start of the period, before the shock is realized:

$$V_{it}(b_{it}, h_{it}) = \int_0^{\infty} V_{it}^{\bullet}(b_{it}, h_{it}, \eta) f(\eta|Z_{it}) d\eta. \tag{11}$$

The retired life cycle problem can thus be solved backwards from the terminal period to retirement age by alternating applications of (10) and (11). Details of the solution method and optimality conditions are omitted here, but are effectively identical to the previous chapter. The only caveat is that the value functions for all periods implicitly depend on the current price of care p_t and the individual's belief about inflation $\hat{\pi}$. That is, the value functions are subjective expectations of the present discounted value of utility, based on current information and belief. When individuals' beliefs are incongruous with reality, they will formulate new beliefs about future medical prices each period as they receive a surprised about the current price of care, and thus must repeatedly re-solve their lifecycle problem.

To summarize results from the previous chapter, the model was estimated using data from the Health and Retirement Study (HRS), a panel study of older Americans that tracks income, assets, health, and medical expenses. Parameter estimates, with brief descriptions of each parameter, are presented in Table 1. All individuals in the estimation sample were assigned to one of 150 "types" depending on their cohort, sex, and income quintile. Each type was assigned an exogenous timepath of income y_{it}, insurance premiums z_{it}, and copay rate q_{it} based on the HRS data, as described in the previous chapter. Income is assumed to be non-stochastic for simplicity, as income risk is greatly reduced for retired individuals.

3 Counterfactual Simulations

The model described above is used to simulate counterfactual scenarios to understand the total effect of medical inflation on consumption, saving, and medical spending, and the decomposition of these changes between the substitution and anticipatory effects. The main counterfactual simulation addresses the dynamics of individuals' choices under historically high medical inflation; additional simulations examine the same under low inflation, and a static decomposition of the own price elasticity of medical care and cross price elasticity of consumption is also presented.

All simulations begin in 2010 with an HRS subpopulation of 694 single, retired Americans between the age of 65 and 70 who have signed up for Medicare and whose total net real assets in year 2000 dollars are between -$3,000 and $8,000,000. This includes 192 men and 502 women, each replicated 100,000 times to allow a large diversity of shocks to occur in each period. Each individual in the simulation population is assigned to one of thirty "types" based on their sex, two-year birth cohort, and income quintile (as determined by the relative ranking of their income). Survival-conditional

timepaths of income are assigned to each type as described in the previous chapter. For each simulation, the 2010 asset and health values (again, constructed identically to the previous chapter) are used as initial conditions; the sample is replicated 100,000 times to allow each HRS-observed individual to receive a broad array of shocks to health and medical needs over time.

While the structural model was estimated in the previous chapter using single retired Americans born between 1910 and 1939 in the HRS dataset, the counterfactuals of this chapter use individuals born between 1940 and 1945. The estimation required that any included cohort be present for at least three data waves after being on Medicare, thus only individuals who had reached the eligibility age of 65 by 2004 could be used in the estimation. To examine the dynamic effects of medical inflation on the elderly over the course of their retirement, it is best to start with younger retirees. This maximizes the number of periods that can be simulated, as attrition due to death is high. Moreover, it allows the simulated population to be relatively homogeneous, so that the time dimension can be roughly reinterpreted as age. New cohorts are not added to the simulation as they reach retirement age, as this paper only examines the effects of medical inflation on the already retired. Moreover, the model is not equipped for the working population, so it cannot predict how changes in inflation will affect the savings decisions of the working population as they enter retirement, nor the timing of their retirement.

The baseline scenario against which all counterfactual simulations are compared is defined by all individuals believing that medical inflation will be constant at its historical average of 2.24%, and this rate actually does occur ($\pi = \hat{\pi} = 0.0453$ with two year periods). In a simulation, individuals' states and choices are recorded in each period and then aggregated for reporting. For example, the consumption levels of all individuals in the fifth period of the simulation are tracked, and the sum of these values is reported. All counterfactual figures are displayed in terms of percentage deviation from the baseline: counterfactual divided by baseline, less one. In all cases, the overall shape or trajectory of the object of interest is unchanged between simulations, so this "de-trended" format does not obscure any important features, but better elucidates the differences between scenarios, particularly the decomposition of the changes between the intra- and inter-temporal effects. Some of the objects of interest are decomposed by income and wealth quintiles for a period midway through the simulations and presented in Tables 2-6; these will be discussed more fully in their respective sections. A welfare analysis is also presented, decomposed by income and wealth.

In any given period, the aggregated simulation data includes only survivors. Over time, the pool of survivors becomes wealthier in terms of both income and assets relative to the initial sample,

as these individuals begin the simulation with higher health on average. Unfortunately, there is no easy way to correct for this bias, as death cannot be "turned off" in the simulations without consequences that likely outweigh the benefits of eliminating the survivor bias. Because health is a major determinant of the magnitude of both medical needs shocks and mortality shocks, artificially preventing simulated death would generate a severely sicker population, and thus much higher medical spending. Moreover, it is not clear what to do with individuals that would die from adverse health shocks rather than mortality shocks (in (8), those with $\hat{h}_{it+1} \leq 0$): should the shocks be redrawn, or should the individual be assigned to live with $h_{it+1} = 0$, or simply carry on living with negative health? Luckily, most effects of the counterfactual simulations do not vary considerably by assets and income, except among the richest and poorest groups. Likewise, changing the prices of medical care alters the level of health investment purchased by simulated individuals, and thus their future health and survival prospects. This effect is actually very small, so that there is little difference in survivor bias between the decomposed elements of a counterfactual.

Insurance premiums and coinsurance rates are treated as exogenous functions of the individual's age, sex, health, and income, and interactions thereof. That is, each individual is assigned a reduced form insurance contract based on the average contract purchased by individuals with the same characteristics. These insurance functions were estimated in the previous chapter based on data in the HRS between 1996 and 2010 for individuals in the structural estimation sample and did not control for year and thus implicitly omitted the price of medical care. As the parameters of the model were selected to match various features of the HRS data while assuming that premiums would not steadily rise in the future due to medical inflation (changing individuals' expectation about their need to retain assets), the current simulations take the same approach. Insurance premiums do not represent a large portion of spending by retired individuals, as most insurance is acquired heavily subsidized through Medicare. Thus omitted growth in insurance premiums will not severely alter the timepaths of the variables of interest.

The dynamic simulations of high and low medical inflation (the "alternate rate") decompose the effects with three alternative simulations. First, the total effect of the change in medical inflation, including both current period effects of the price of care and anticipatory effects of future higher prices. This occurs when the alternate level of medical inflation (different from the historical average) occurs in each period and individuals believe that this alternate level will arise. Second, the intraperiod or reactionary effect, which is captured by the alternate inflation rate actually occurring while individuals believe that the baseline rate will occur. Third, the interperiod or precautionary

effect, which arises when the historical baseline rate of inflation occurs, but individuals believe that the alternate rate will occur. The second and third simulations form a nearly perfect decomposition of the first simulation: the sum of the "reactionary effect" and "precautionary effect" trajectories equals the "total effect" trajectory in Figures 3-14 at all periods of life. This can be visually confirmed on any of the figures by noting that whenever the red or green line crosses the horizontal axis, the other color crosses the blue line at the same time.[4] There seems to be very little interaction between the two effects, allowing a clear analysis of the impetus for behavioral changes.

Individuals in the precautionary effect and reactionary effect simulations always see the current period price of care and act on it optimally, but their beliefs about the future diverge from reality; they are surprised each period by the price of care, as it was not the one they had anticipated. In terms of the model and behavior, this inconsistency requires re-solving the life cycle each period. When a new period begins and the individual learns the actual price of health care in that period, he generates new beliefs about the stream of future prices $\hat{p}_{ts}$ according to (2). As different choices would be made in the future under these prices, behavior in all remaining periods must be recursively re-optimized from the terminal age as described in the previous section. The computational burden of having inconsistent beliefs is thus on the order of T^2, where T is the total number of periods from retirement to the terminal age.[5] Beyond the standard benefits of restricting analysis to retired individuals (breaking the causal link from health to income, less uncertain income, etc), this exercise is only possible when T is small enough for T^2 to be a reasonable number. With individuals' lives capped at 105 years and using two-year periods (as in the HRS data) from age 65, a total of 231 periods must be solved for each type, a large but not infeasible problem.

Analysis of the dynamic effects of medical inflation will focus on the high medical inflation simulation, in which the alternate inflation rate is 3.51%. This corresponds to the average rate of medical inflation from 1981-1993, with a baseline historical rate of 2.24%; the results of these simulations are presented graphically in Figures 3-8.[6] The analysis will proceed as follows: Sections 3.1, 3.2, and 3.3 will discuss the dynamic effects of high medical inflation on an array of outcomes aggregated across individuals, Section 3.4 decomposes these changes by income and wealth subgroups and considers a welfare analysis, Section 3.5 briefly discusses and contrasts the dynamic effects of low medical inflation, and Section 3.6 presents a static decomposition of the price elasticity of medical

[4]The near perfect additivity breaks down somewhat at very late ages, but is still very close.

[5]The actual number of periods to be solved is close to $\frac{1}{2}T^2$.

[6]Note that in all of these figures (and those for the low inflation counterfactual), the values for all three simulations are equal to the baseline in the first period, as no change has yet occurred and the individuals are still in the baseline world.

care.

3.1 Assets and Spending

To begin the analysis, Figures 3 and 4 demonstrate the effects of higher medical inflation on asset holdings and total spending. Here, total spending is defined as the sum of all individuals' out-of-pocket payments on consumption, medical consumption, and health investment:

$$\text{Total spending} = \sum_i \left(c_{it} + p_t q_{it}(\kappa_{it} + \mu_{it}) - w_{it}\right). \tag{12}$$

Note that welfare payments from the government (to meet the utility floor in extreme circumstances) are netted out of spending so that the object represents only out-of-pocket spending on the three goods. Insurance premiums z_{it} are also not included, as they are assumed to be exogenous and would only shift the scale of the figure after normalizing by baseline spending.

In the period in which individuals learn that inflation will be perpetually higher, they cut spending by 0.4% (see Figure 4, "total effect" blue line). The price of care is identical to the baseline scenario, but individuals anticipate future higher prices and seek to retain more assets to pay for the higher health care costs they will face in the future. Once the higher-than-baseline prices actually arrive, the lower spending level persists for about ten years, but eventually rises to overtake baseline spending sixteen years after the higher rate took hold. Total spending continues to rise for the rest of the individuals' lives, eventually exceeding 3% above baseline for the few who survive for thirty years.

As expected, the initial reduction in spending allows retired individuals to accumulate assets faster than baseline (or decumulate them slower), having about 3.3% greater wealth by the time they reach the "break even" spending point after sixteen years (Figure 3. However, even once total spending begins to rapidly accelerate later in life, assets continue to grow relative to baseline. That is, the additional assets accumulated in the first sixteen years are sufficient to sustain and grow themselves even when spending occurs at a significantly higher rate. Rather than accumulate or hold assets for later use (winding them back down toward the baseline level), precautionary saving continues indefinitely. The fact that retired individuals would find it optimal to continue to hold higher levels of assets reinforces how strong a role uncertain medical expenses play in motivating savings behavior late in life.

Decomposing these changes by the intratemporal ("reactionary effect" red line) and intertem-

poral ("precautionary effect" green line) effects, note that total spending in the reactionary effect simulation is approximately a lagged version of the total effect simulation, with no initial sudden drop in spending. Individuals in this simulation do not take any precautionary measures against future price increases, thus their total spending only begins to fall once the higher prices for medical care begin to arrive. Even then, spending falls much more slowly than under the total effect simulation, as the retirees only partially foresee the even higher prices in the future. That is, these individuals know that the price of medical care will be higher than the baseline in the future, but only because the price they *currently* see is already higher, not because they realize that price growth has accelerated. Thus their precautionary saving for high future medical costs is slow to begin, resulting in delayed (and lower) relative asset accumulation. By the time the "total effect" individuals have increased their spending above the baseline, those with unchanged beliefs about medical inflation have only recently bottomed out in their spending decline. Given the relatively high levels of assets attained when medical inflation is correctly observed, it is clear that these individuals have suboptimally low savings later in life. Note that this imprudence due to a lack of foresight (or adaptive beliefs) will only come back to sting an individual if he would live sufficiently long to experience the lower spending levels at high ages; individuals in the reactionary effect simulation who die early in their retirement get the pleasures of their imprudent spending but never have to experience reduced consumption later in life due to a lack of assets.

In contrast, individuals in the precautionary effect simulation consistently over-save in anticipation of higher future medical inflation that never actually arrives. When these individuals first change their beliefs about the rate of medical inflation, they cut spending by the same 0.4% as if they were in the total change scenario. However, after advancing one period and finding that the price of medical care is not as high as expected, these individuals can now spend more than their total effect counterparts both because of the lower current price and because of the downward revision in expected future prices. In subsequent periods, additional assets have been retained relative to baseline, providing additional ability to increase spending. In this way, the trajectory of relative asset accumulation tracks very closely with the total effect simulation in early periods, but rapidly diverges as the precautionary effect individuals increase their spending much more rapidly. While these individuals are able to continuously expand their spending, exceeding baseline levels barely ten years after the change in beliefs, they do so at a very moderate rate. Even very late in life, when there are very few possible periods left in which medical inflation can adversely affect survivors, individuals continue to accumulate assets above baseline. In summary, medical inflation-motivated

asset accumulation is initially driven by foresight or belief in future higher prices, but later in life the signal value of actually higher prices is the primary impetus for continued saving.

3.2 Consumption and Medical Costs

Turning attention from overall spending to the composition of behavioral changes due to higher medical inflation, Figures 5 and 6 demonstrate the paths of consumption and out-of-pocket medical spending for the three simulations, as compared to the baseline scenario. So that the two components shown in these graphs add up to the total spending of the previous figure, out-of-pocket medical expenses are calculated net of any welfare payments received: $OOP_{it} = p_t q_{it}(\kappa_{it} + \mu_{it}) - w_{it}$. Again beginning with the total change simulation (in blue), Figure 5 reveals that higher medical inflation causes an immediate cut of about 0.4% in consumption as individuals undertake more precautionary saving. As the price of medical care will be higher next period (and all subsequent periods), there is an income effect as an individual cannot attain as much utility with any given level of expenditure. The expected lifetime value of arriving at any given state (a b_{it}, h_{it} pair) is lower than the baseline, while the marginal value of assets is higher. To satisfy the first order conditions that the marginal utility of consumption must equal the (discounted) expected marginal value of wealth next period, this induces lower consumption in the present. Consumption is persistently lower for over twenty years, steadily between 0.4% and 0.5%, until turning upward later in life and eventually exceeding the baseline in the final few years as their large reserve of additional wealth allows them to finance more consumption. In short, individuals cut consumption initially, but then maintain nearly the same rate of consumption growth as in the baseline scenario.

The trajectory of out-of-pocket medical expenses is much different (see Figure 6). After an initial drop (but by a greater amount than consumption, about 0.7%), medical expenditures rise rapidly, passing the baseline level within four years and exceeding 10% above baseline by the time the simulated individuals are in their early 80s. Medical spending continues to rise for the rest of the individuals lives, only slowing in growth at ages approaching 100. Unsurprisingly, the higher prices attained in the high medical inflation scenario dominate any precautionary motive: reductions in medical care cannot keep up with rapid inflation. In this way, nearly all of the additional asset holding is due to the consistently depressed level of consumption, while the spending growth seen in Figure 4 is entirely due to increased medical expenses. Note that while medical expenses grow rapidly relative to baseline as soon as inflation increases, total spending remains steadily below baseline for

over ten years. This seeming contradiction is due to the increasing magnitude of out-of-pocket costs as the individuals age, encompassing an ever greater share of total spending.

The decompositions of consumption and medical expenditures by intra- and intertemporal effects likewise show strikingly different patterns. The trajectory of consumption in the precautionary effect scenario is very similar to total spending. After an initial cut in preparation for lower future consumption due to higher medical costs, consumption steadily rises for the rest of the individuals lives, surpassing baseline levels in about ten years. The additional assets saved in early periods are used to finance a higher standard of living, overcoming the precautionary effect of higher anticipated prices for medical care. Individuals in the reactionary effect simulation, however, are repeatedly surprised by higher than expected prices and thus find themselves perpetually unprepared: consumption falls steadily relative to the baseline until individuals reach their early 90s. The decline in consumption occurs despite the individual holding more assets than the baseline, always needing to further reduce his standard of living because his previous cuts were insufficient. In this way, the constant proportional reduction in consumption for individuals in the total effect simulation reflects a balanced tension between preparing for future anticipated price increases and trying to manage those prices as they actually arrive.

In stark contrast, the path of out-of-pocket medical expenses in the reactionary effect simulation tracks very closely with the total effect simulation's rapid increase, while the precautionary effect simulation hews fairly close to the baseline level. Individuals experiencing only the reactionary effect do not initially reduce their purchases of goods (as they do not anticipate additional future need), so their medical expenses are slightly higher for the first few periods once the higher rate of medical inflation arrives. After about ten years, medical expenses for reactionary effect individuals fall below the total effect path due to the lower level of assets retained to finance purchases. Individuals in the precautionary effect simulation, however, do not experience drastic changes in medical expenses. Instead, their relative change in out-of-pocket costs closely matches the pattern for consumption, with an initial cut followed by a steady gradual rise as accumulated wealth allows for larger purchases. This follows from the optimality condition for consumption and medical consumption, which dictates an optimal ratio conditional on price and the medical needs shock. Because medical consumption represents the lion's share of total medical purchases, and there are no price differences (versus baseline) in this simulation, it makes sense for aggregated out-of-pocket medical costs to mirror consumption.

3.3 Medical Consumption and Health Investment

The two components of medical expenses– medical consumption and health investment– also demonstrate different patterns when individuals face higher medical inflation, as shown in Figures 7 and 8. As medical consumption constitutes the vast majority of total medical care purchases $M_{it} = \mu_{it} + \kappa_{it}$, Figure 7 can be viewed as showing the paths of both total medical expenses and (approximately) medical consumption. Even as individuals' medical costs are rapidly increasing through higher prices, this effect is tempered by continuously reducing the quantity of care. After the initial cut of 0.7%, identical to the out-of-pocket cost cut, medical care rapidly decreases for twenty-five years, eventually falling by nearly 9% relative to baseline; after age ninety, individuals have retained enough additional wealth through reduced consumption to stabilize medical care and begin to return it to the baseline level. Just as with out-of-pocket medical expenses, nearly all of the change in the quantity of medical care is due to the reactionary effect of higher prices of medical care actually arriving, rather than precautionary effects from foreseen inflation: the red path tracks very closely with the blue total effect path. Meanwhile, the relative trajectory of medical care for individuals in the precautionary effect simulation (green) stays even closer to baseline levels than does its out-of-pocket counterpart, rapidly stabilizing at about 0.5% after the initial drop. Figures 6 and 7 jointly show how individuals split the burden of higher prices between lower quantities and higher costs, initially favoring the former before shifting the bulk of the burden to the latter. Moreover, when Figure 7 is interpreted as the relative path of medical consumption, it is clear that any utility loss from failing to react to higher medical inflation (as in the reactionary effect simulation) comes from reductions in consumption rather than medical consumption, as there is very little difference in quantity whether or not the higher inflation is acknowledged.

The relative path of health investment when individuals face higher medical inflation is similar in shape to that of medical consumption, but on a much greater scale. While medical consumption levels out at a 9% reduction in quantity after about twenty-five years, purchases of health investment continue to fall for individuals' entire life, at a rate nearly three times as fast. The estimated model thus reveals that health investment is a very price elastic good. This is consistent with the shape of the health production function shown in the previous chapter, which is initially very steep before quickly leveling off once moderate quantities of health investment have been purchased. Because most individuals will choose a level on the highly curved portion of the production function, even small changes in the cost of additional health will elicit a large quantity response to match

the marginal benefit of higher future health. The other salient difference between the trajectories of the two medical care goods is the faster and perpetual increase in health investment in the precautionary effect scenario, which eventually approaches 2.5% above baseline at the very end of life– even greater than the increase in consumption. This indicates that health investment is also significantly more sensitive to wealth than the other goods, and is in fact a superior good. This is again consistent with the estimation of the previous chapter, which found significant differences in the rate of health decline by wealth, holding income constant, indicating greater health investment with larger wealth. The Grossman model of investment in a capital stock of health predicts greater investment when price growth is faster: individuals substitute health investments from the future and to the present because of the change in the intertemporal price ratio. This effect seems to be small in this model, or at least overwhelmed by the precautionary motive that induces the initial cut. The intertemporal substitution effect predicts an initial increase in the level of health investment when the individual learns of higher medical inflation, but no effect on its subsequent growth rate. In fact, the precautionary effect simulation reveals a *greater* initial drop in health investment than the other two goods, followed by faster growth. In this way, Grossman's prediction does not hold when individuals make endogenous savings decisions.

3.4 Varying Effects by Income and Wealth

The analysis above aggregates all individuals and reports the path of population averages relative to the baseline level. However, not all individuals respond in the same way, as there is considerable variation in income and wealth in the simulation sample. While any reader would find it tedious to examine repeated copies of the six simulation figures discussed above, broken out by income and wealth, these differences can be summarized at a single point in time. To this end, Tables 2, 3, 4, 5, and 6 present changes in several variables of interest fourteen years after the onset of higher inflation, decomposed by income quintile (in rows) and wealth quintile *within* each income quintile at the start of the simulation (by columns); the rightmost column aggregates all individuals within an income quintile. The values presented are the "total change" percentages that would appear in Figures 3, 5, 6, 7, and 8 at year 14 if those simulations were restricted to only that particular subpopulation. This timing was selected because it is approximately half way through the entire simulation (at least for those with long lifespans), and is fairly close to life expectancy at the start of the simulation. These tables are presented to gauge the relative magnitudes of the effects already

discussed, rather than provide precise values; the subpopulations are based on rather small segments of individuals in the HRS data. Because of the aggregation, the "All" column tends to be strongly weighted toward the fourth and fifth wealth quintiles, where all values are much larger in absolute level.

Beginning with relative changes in assets in Table 2, the most salient pattern is that individuals with greater wealth within each income quintile hold less additional assets to self-insure against high future expenses. This is sensible, as the wealthiest individuals are already well prepared for high medical inflation and thus need not retain even more assets. Among the lower three wealth quintiles, individuals retain relatively fewer additional assets at higher income quintiles, likely for a similar reason. Higher income individuals have greater assets within any given wealth quintile,[7] and are better able to finance higher costs from their future income stream. The trend is incoherent for the fourth wealth quintile, and is mildly reversed for the wealthiest groups. The differential effect on consumption is significantly flatter (see Table 3), but with a slight trend for larger cuts among richer individuals. While the rich have greater resources and are better protected against high medical costs, even small changes in consumption among the poor generates considerable utility loss due to their much greater marginal utility.

Table 4 reveals that relatives changes in out-of-pocket medical expenses are fairly flat across income quintiles overall, but with different patterns by wealth level within each income quintile. Among the lowest income quintile, out-of-pocket costs show larger growth for individuals with more wealth. At first blush, one might expect that wealthier individuals are using their assets to pay the higher prices and avoid cutting their purchases of care, but this is not the case: the first row of Table 5 reveals that the reduction in medical care quantity is also larger with higher wealth. How can individuals cut quantity more but still experience a greater increase in cost? The answer to this paradox lies in the structure of insurance in the model. Insurance coinsurance rates were estimated as a function of age, sex, health, income, and their interactions; health was found to be positively correlated with coinsurance, so that insurance is more generous with worse health. Among low income individuals, health is positively correlated with wealth, so those with more assets pay a greater portion of their medical costs out-of-pocket and thus must bear more of burden in both cost and reduced quantity. Moreover, the lower wealth quintiles of the bottom income quintile are very similar in terms of assets, with everyone very close to zero. These individuals are constrained by the

[7]Moving vertically in these tables shifts assets as well as income, as the Nth wealth quintile of the third income quintile has greater assets on average than the Nth wealth quintile of the second income quintile. Comparisons across income levels should thus be interpreted with caution.

utility floor much more frequently, requiring welfare payments to achieve the minimum standard of living while maintaining non-negative assets. As medical costs become even greater with medical inflation, these individuals are protected from higher costs because they simply *cannot* pay any more. In contrast to the poorest, the highest income quintile experiences smaller relative changes in medical expenses with greater wealth. This is due to variation in the composition of medical care by wealth, described fully in the next paragraph. The upward pattern is also present for individuals in the second income quintile, but at a greatly reduced scale; the third and fourth income quintiles are flat or show a slight inverted U-shape. The middle income quintiles thus smoothly transition between the two different effects as they trade dominance.

Turning to the quantity of medical care, Table 5 shows a very clear trend: richer individuals reduce their purchases of medical care by more than poorer individuals, whether examined in terms of income or wealth. Again, this seems to run counter to intuition, as these individuals are better prepared to deal with the difficulties of higher medical prices and should thus need to change their behavior less to adapt. The pattern is explained by the fact that health investment is a superior good, considered in terms of both income and assets, and thus makes up a larger share of medical care for richer individuals. As described above, health investment is also more sensitive to price, with its relative quantity rapidly falling throughout the high inflation simulation. The aggregate pattern holds for all subgroups,[8] but the larger share for richer individuals generates a greater decline in total medical purchases relative to baseline. Note that in each income quintile row, the trend accelerates as one moves from lower to higher wealth quintiles; this matches the highly skewed, top-heavy distribution of assets among retired individuals. A similar accelerating trend is present when considering increases in income, both in aggregate and within relative wealth ranking.

Finally, Table 7 presents the aggregate equivalent variation (EV) of the increase in medical inflation across income and wealth subgroups, as a percentage of assets at the beginning of the simulation (2010 HRS data). As the solution method presented in Section 2.3 tracks individuals' value functions in each period of life, it is possible to calculate for each individual the cash equivalent of facing faster price growth: the amount of assets that would make him indifferent between losing the money and maintaining baseline inflation *or* keeping the money and experiencing high medical inflation. Within each subgroup, the reported percentage is the sum of each individual's EV divided by the sum of their assets, for consistency with the methods used in previous tables and to prevent strange results

[8] See Table 6, which shows that the relative decline in health investment is nearly identical across all individuals, except for the very poor.

among the poorest groups. Overall, higher medical inflation is equivalent to losing about 2.25% of wealth, with very little difference by income level. Disaggregated by wealth within each income quintile, the relative losses nearly universally decline with greater wealth. Even though consumption falls by slightly less among those with fewer assets, the utility loss represents a significantly greater share of wealth due to the concavity of the utility function (higher expected marginal utility of consumption among the poorer groups). Moreover, the denominator of total wealth is much smaller for these groups (near zero for the absolute poorest), allowing even relatively small absolute cash values to be larger percentages of aggregate wealth. While the utility losses are not trivial, EV analysis reveals that a return to the high medical inflation of the 80s and early 90s would be costly but not crippling for most retired individuals.

3.5 Low Medical Inflation

While the analysis above focused on the dynamic effects of historically high medical inflation, it is also worth considering how significantly lower inflation, as occurred from 1994-2000, would affect retired individuals. To this end, Figures 9-14 graphically present the aggregate paths of variables (and their decompositions) of interest discussed in the preceding sections, with an alternate medical inflation rate of 1.09% and the same baseline rate of 2.24%. A cursory examination of these figures demonstrates that most of the effects of changes in medical inflation are approximately linear, even in a dynamic setting: they are near mirror images of Figures 3-8, reflected across the horizontal axis and with a shifted scale. While a complete analysis of low inflation scenario simulations would be redundant and tedious, a brief discussion is warranted to give a flavor of how the effects of inflation are inverted and highlight key differences when applicable.

When individuals become aware that medical inflation will be perpetually lower, they respond by immediately increasing spending on all three goods in anticipation of their greater future purchasing power and lower marginal value of wealth. Their greater spending depletes their asset holdings faster than in the baseline scenario. Just as with high inflation, this effect lags behind for individuals in the reactionary effect simulation, who do not perceive lower price growth and only respond to price changes as they occur. As time passes and the lower medical prices are realized, out-of-pocket costs of care fall rapidly relative to baseline, significantly reducing total spending later in life. However, aggressive purchasing early on overwhelms these savings, and assets continue to fall throughout individuals' lives. The relatively stable elevated level of consumption represents a tension between

individuals' eagerly anticipating lower future prices (and thus rapidly increasing consumption before gradually cutting it due to lower wealth) and a slower response to the current price as a signal of future costs.[9] Oddly enough, prudent foresight dictates "living it up" in the present when faced with lower medical inflation.

While the shape of the trajectories of the variables are generally the same (but inverted) as in the high inflation scenario, the scale of these changes is not. With a baseline inflation rate of 2.24%, the high rate is 1.27% greater, while the low rate is 1.15% smaller; the former thus represents a more significant change in inflation by a factor of about 1.1. Accordingly, we see that the inverted relative paths of assets, consumption, etc in the low inflation scenario are scaled down by about 10% from their high inflation counterparts at each period of life. In this way, the aggregate effects of permanent changes in medical inflation are relatively linear in the magnitude of the change, even in a dynamic setting in which behavioral decisions are partially motivated by changes in state variables due to previous decisions. More surprisingly, this aggregate linearity emerges from a model that is highly non-linear in its fundamental structures and individual behavior.

The only variable of interest whose trajectory does not scale down with the smaller change in inflation is health investment. Rather than increasing about 10% slower than the rate of relative decrease for individuals in the high inflation scenario, the lower inflation yields a relative growth rate about 20% faster in all periods. As noted above, health investment is a superior good in the model, with wealthier individuals (in terms of both income and assets) dedicating a greater portion of their spending to health investment. With slower price growth for medical care, all individuals experience a significant income effect as they are now able to purchase greater quantities of goods than before. These benefits are taken disproportionately in the form of increased health investment: additional periods of life are now more attractive due to the higher levels of consumption that will be attained, thus individuals seek to stay healthier to increase longevity. The non-linearity in health investment versus medical inflation results from the extensive margin as well: some poorer or sicker individuals who previously did not invest in their health will make non-zero choices under lower medical inflation.

[9]The additivity of the reactionary effect and precautionary effect trajectories to nearly perfectly yield the total effect path is somewhat weaker under low inflation. In some of the figures, visual inspection reveals noticeable deviations, particularly at very high ages. This discrepancy is not further explored here.

3.6 Static Decomposition

In the dynamic simulations, the reactionary effect simulation tracked very closely with the total effect simulation for all variables concerning medical care. That is, the current price of medical care and its signal of future prices was the main determinant of medical care behavior, whereas knowledge of changes in the rate of inflation had a comparatively negligible effect. As an alternative analysis of demand for medical care, I perform a static (one period) decomposition of the elasticity of demand, splitting the effect of the current period price of medical care from its signal of future prices; the results are presented graphically in Figures 15 and 16.

Individuals in the "total change" simulation (TC, blue line) believe in a constant baseline rate of inflation, but the initial price of medical care is altered to various levels within $\pm 5\%$ of the true 2010 price. The variable of interest is the percent change in total quantity of medical care purchased in the first period only, relative to quantity at the baseline price. Individuals in this simulation interpret the change in price as both a change in their current budget constraint and as an indicator of future prices they will face, altering their marginal value of wealth. The slope of the blue demand curve represents the full single period price elasticity of demand for medical care, shown to be about -0.462 on average over this range. Demand for medical care is thus moderately inelastic, consistent with many previous estimates based on observational data from actual price changes, though somewhat above the median estimate (in absolute value). Even though the estimation of the structural model in the previous chapter is not explicitly based on variation in the price of care, the simulation still yields an acceptable elasticity of demand for care, further validating the model.

To divide the effect of current price from its signal value, the red and green demand curves assign different beliefs about inflation to the simulated individuals. For the "present change" simulation (PC, red line), the price deviations *only* affect the current period price of care, as if it were a temporary price shock with no informational value for future prices. These individuals believe that prices will return to the same price path as in the total change simulation in all subsequent periods. In contrast, individuals in the "future change" simulation (FC, green line) experience no change in price in the period of record, but believe that all future prices will follow the path of the total change simulation. Defining a as the percent adjustment to the initial price of care (which takes values from 0.95 to 1.05), and p_0 as the baseline initial price, three price paths are:

$$\vec{p}_{TC} = \langle ap_0, a\pi p_0, a\pi^2 p_0, ... \rangle, \quad \vec{p}_{PC} = \langle ap_0, \pi p_0, \pi^2 p_0, ... \rangle, \quad \vec{p}_{FC} = \langle p_0, a\pi p_0, a\pi^2 p_0, ... \rangle. \tag{13}$$

As seen in Figure 15, these alternate simulations do indeed decompose the total effect of changes in the price of medical care, as the red and green lines sum nearly perfectly to match the blue line. The complete lack of an interaction effect between present and future price changes is unsurprising given that a similar decomposition was consistent even in a long dynamic simulation, when state variables evolve differently over time. Saliently, nearly all of the price elasticity of medical care is due to the current price, rather than what it signals about future prices: elasticity in the present change simulation is -0.434, while it is only -0.027 for the future change simulation. That is, the future signal value from changes in price accounts for only 6% of the change in quantity, while the current price itself accounts for the remaining 94%. This is a startling result, revealing that individuals put far more weight on one single period in the present than all future periods combined. Individuals do not choose to smooth their purchases of medical care when they anticipate a permanent change in price one period in the future. The result from the dynamic simulations is thus strengthened: nearly all of the sensitivity of medical care to price is due to present effects, rather than any consideration for future price level or growth rate. Moreover, almost all precautionary behavior comes in the form of adjustments to consumption.

In contrast, the decomposition of the cross price elasticity of demand for consumption with medical price is shown in Figure 16. Unsurprisingly, consumption is much less sensitive to the price of medical care, with an elasticity less than one twentieth the own price elasticity of care ($\epsilon = -0.023$). More interesting is the reversal of the source of this sensitivity: rather than reacting to changes in the current price as with medical care, changes in consumption occur almost entirely because a higher price in the present signals a stream of higher prices in the future. In fact, the proportions are nearly exactly reversed, with the "future change" portion accounting for 94% of cross price elasticity and "present change" only 6%. This confirms the findings of the dynamic simulations, in which cuts to consumption were the primary mechanism by which individuals saved more wealth to pay for higher future medical costs. While one period changes in the price of medical care are treated as mere transitory shocks that have nearly no effect on consumption, permanently higher future prices drastically affect the marginal value of wealth and induce lower consumption immediately.

4 Conclusion

This chapter extends the previous chapter's structural model of retired individuals' decisions about consumption, medical care, and saving to explore the dynamic effects of changes in medical inflation. In particular, it focuses on differentiating between two aspects of inflation: first, the intratemporal or reactionary effect from individuals experiencing higher prices of care as they occur; and second, the intertemporal or precautionary effect from individuals foreseeing future price growth. Using a replicated sample of younger retirees from the 2010 wave of the Health and Retirement Study, the previously estimated model is used to simulate their remaining lifespans under different combinations of actual medical inflation and beliefs about inflation. The total effect of individuals correctly perceiving a permanent change in inflation just before it begins is decomposed into a simulation in which inflation has changed but individuals believe it actually has not (reactionary or intratemporal effect), and a simulation in which inflation has not changed but individuals believe it has (precautionary or intertemporal effect).

The counterfactual simulations reveal that higher medical inflation would induce retirees to retain more assets for the rest of their lives, running down their savings more slowly. The greater uncertainty that arises from higher variance of medical costs motivates individuals to hold more precautionary savings, more than offsetting the downward pressure on assets due to the increase in costs. The increase in both savings and medical expenses are financed by a permanent cut to consumption, a constant proportion of the baseline under the historical average rate of inflation. Even as medical costs skyrocket, retirees also deal with higher medical inflation by cutting the quantity of care purchased. While the rate of increase of medical expenses is similar to the rate of decrease for the quantity of care, the former continues to grow while the latter stabilizes later in life once individuals achieve sufficient wealth. However, the two medical goods in the model– medical consumption and health investment– do not follow the same trajectory. Health investment is much more price elastic than medical consumption and its quantity purchased falls about three times faster, never stabilizing with greater wealth. This occurs because not only has the cost of health investment increased, but its returns have decreased: lower future consumption makes additional healthy periods of life less attractive. Moreover, the classical result that faster price growth causes higher health investment through intertemporal substitution does not hold in an environment with endogenous savings, or at least is a very small effect.

The decomposition of these effects into precautionary and reactionary portions reveals starkly

different patterns across goods. The steadily lower level of consumption is generated by a balanced tension between the two effects. Foresight of higher future prices increases the marginal value of wealth, leading to an immediate precautionary cut in consumption; as individuals accrue additional assets but the higher inflation doesn't actually arrive, individuals steadily increase their consumption relative to baseline. On the other hand, individuals' reaction to higher medical prices as they arrive causes consumption to continuously fall relative to baseline levels as individuals find themselves perpetually unprepared for high medical costs. They learn that future prices will be higher than previously expected, but only because the price is already higher in the present; by never anticipating faster price growth, the reactionary effect is always playing catch-up. In contrast, nearly all of the change in medical expenses and quantity of care (of both medical goods) is attributable to the reactionary effect as higher prices are actually realized. While the precautionary effect on medical care follows a similar pattern as for consumption, it is overwhelmed in magnitude by the rapid growth in medical prices that generates greater expenses even as quantities are cut. Moreover, a static decomposition of the price elasticity of medical care shows that this is truly an intratemporal effect from the higher current price itself, rather than a result of the signal it carries about future prices. This is reversed for the cross price elasticity of consumption, where sensitivity to medical prices changes depends almost entirely on the signal value.

Beyond pure economic interest, these results provide optimistic evidence that dynamic general equilibrium models of insurance contract choice with endogenous premiums would not be cripplingly difficult to solve and estimate. When both consumers' choice of insurance contract and the menu of contracts offered are endogenous choices to be found in equilibrium, individuals' valuation of contracts will depend on their beliefs about the contracts that will be offered to them in future periods. Those future contracts in turn depend on current period choices and outcomes, creating an intractible rational expectations problem. As this chapter demonstrates that individuals' demand for medical care depends almost entirely on current conditions rather than future expectations (at least with similarly specified preferences), it might not be necessary to painstakingly seek perfect consistency between current market outcomes, future conditions, and beliefs about the future. If agents choices in the present are highly inelastic with respect to changes in future market conditions (through medical prices or coinsurance rates), then an approximation of the true future is sufficient to accurately solve for current behavior.

References

().

Aizcorbe, A. and Nestoriak, N. (2011). "Changing Mix of Medical Care Services: Stylized Facts and Implications for Price Indexes." *Journal of Health Economics*, 30(3): 568–574.

Berndt, E. R., Cutler, D. M., Frank, R. G., Griliches, Z., Newhouse, J. P., and Tripplett, J. E. (). *Handbook of Health Economics*, vol. 1, chap. 3, pp. 119–180. Elsevier.

Bundorf, M. K., Royalty, A., and Baker, L. C. (2009). "Health Care Cost Growth Among The Privately Insured." *Health Affairs*, 29(5): 1294–1304.

Carroll, C. D. (1997). "Buffer Stock Saving and the Life Cycle/Permanent Income Hypothesis." *Quarterly Journal of Economics*, 112(1): 1–56.

Carroll, C. D. and Samwick, A. A. (1997). "The Nature of Precautionary Wealth." *Journal of Monetary Economics*, 40(1): 41–71.

Dunn, A. (2013). "Health Insurance and the Demand for Medical Care: Instrumental Variable Estimates using Health Insurer Claims." *Unpublished working paper*.

Dunn, A., Liebman, E., Pack, S., and Shapiro, A. H. (2012). "Medical Care Price Indexes for Patients with Employer-Provided Insurance: Nationally Representative Estimates from MarketScan Data." *Health Services Research*, 48(3): 1173–1190.

Dunn, A., Liebman, E., and Shapiro, A. H. (2012). "Developing a Framework for Decomposing Medical-Care Expenditure Growth: Exploring Issues of Representativeness." *Measuring Economic Sustainability and Progress*, forthcoming.

Dunn, A., Liebman, E., and Shapiro, A. H. (2013). "Implications of Utilization Shifts on Medical-Care Price Measurement." *Health Economics*, forthcoming.

Dunn, A., Liebman, E., and Shapiro, A. H. (2014). "Decomposing Medical-Care Expenditure Growth." *Unpublished working paper*.

Elmendorf, D. (2010). "The Long Term Budget Outlook." *CBO Report*, http://www.cbo.gov/ftpdocs/115xx/doc11579/06-30-LTBO.pdf.

Geithner, T. F., Solis, H. L., Sebelius, K., Astrue, M. J., and Frizzera, C. M. (2009). "Annual Report of the Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds." *CMS Report*, https://www.cms.gov/ReportsTrustFunds/downloads/tr2009.pdf.

Levit, K., Smith, C., Cowan, C., Lazenby, H., and Martin, A. (2002). "Inflation Spurs Health Spending In 2000." *Health Affairs*, 21(1): 172–181.

Newhouse, J. P. and Group, R. C. I. E. (1993). *Free for all?: lessons from the RAND health insurance experiment*. Harvard University Press.

Newhouse, J. P. and Phelps, C. E. (). *Handbook of Health Economics*, chap. 7, pp. 261–320. NBER.

Newhouse, J. P. and Phelps, C. E. (1974). "Price and Income Elasticities for Medical Care Services." *RAND Monograph*, R-1197-NC.

Orzsag, P. (2007). "The Long Term Outlook for Health Care Spending." *CBO Report*, http://www.cbo.gov/ftpdocs/87xx/doc8758/11-13-LT-Health.pdf.

Peden, E. and Freeland, M. (1995). "A Historical Analysis of Medical Spending Growth, 1960-1993." *Health Affairs*, 14(2): 235–247.

Ringel, J. S., Hosek, S. D., Vollaard, B. A., and Mahnovski, S. (2002). "The Elasticity of Demand for Health Care: A Review of the Literature and Its Application to the Military Health System." *National Defense Research Institute Monograph*.

Thorpe, K. E. and Howard, D. H. (2006). "The Rise In Spending Among Medicare Beneficiaries: The Role Of Chronic Disease Prevalence And Changes In Treatment Intensity." *Health Affairs*, 25(5): 378–388.

Thorpe, K. E., Florence, C. S., and Joski, P. (2004). "Which Medical Conditions Account For The Rise In Health Care Spending?" *Health Affairs*, W4: 437–445.

Table 1: Model Parameters Used in Simulations

Parameter	Value	Description
δ	0.937	Intertemporal discount factor
ρ	2.06	Coefficient of risk aversion for composite consumption
ν	3.28	Coefficient of risk aversion for medical consumption
ς_0	0.292	Utility level shifter; value of living
α_1	-0.458	Change in marginal utility with health
$\underline{u}$	-4.00	Utility floor (not estimated)
ω_0	1.28	Curvature of bequest motive
ω_1	4.18	Intensity of bequest motive
β_0	-21.96	Base value of medical needs scope parameter
β_s	-2.17	Change in medical needs scope for males
β_{a1}	0.307	Change in medical needs scope with age
β_{a2}	0.00064	Change in medical needs scope with age squared
β_{h1}	2.57	Change in medical needs scope with sickness
β_{h2}	9.36	Change in medical needs scope with sickness squared
β_{k0}	0.108	Base value of medical needs shape parameter
β_{k1}	0.011	Change in medical needs shape with health
γ_0	0.0264	Base health transition level
γ_s	-0.00495	Change in health transition for males
γ_{a1}	-0.00246	Change in health transition with age
γ_{a2}	-0.00001	Change in health transition with age squared
γ_{h1}	0.820	Change in health transition with health
γ_{h2}	0.159	Change in health transition with age squared
γ_{m0}	-12.47	Curvature of health investment production
γ_{m1}	-7.58	Base efficacy of health investment
γ_{m2}	-0.0155	Change in efficacy of health investment with sickness
σ_0	0.0985	Base standard deviation of health transition
σ_1	0.0417	Change in health transition s.d. with sickness
θ_0	-2.69	Base mortality probit level
θ_s	0.452	Change in mortality for males
θ_{a1}	0.0170	Change in mortality with age
θ_{a2}	0.00139	Change in mortality with age squared
θ_{h1}	1.47	Change in mortality with sickness
θ_{h2}	0.608	Change in mortality with sickness squared

Table 2: Difference in Asset Holdings Relative to Baseline by Income and Wealth Quintile, Fourteen Years Into High Medical Inflation Scenario

Income	Wealth Quintile					
Quintile	Lowest	Second	Third	Fourth	Highest	All
Lowest	14.99%	10.96%	7.03%	2.28%	1.58%	1.90%
Second	6.50%	9.42%	6.34%	3.15%	1.57%	2.40%
Third	10.29%	4.22%	3.00%	2.00%	1.90%	2.41%
Fourth	3.27%	3.14%	2.45%	2.26%	2.10%	2.32%
Highest	3.27%	2.38%	2.66%	2.54%	2.41%	2.53%

Table 3: Difference in Consumption Relative to Baseline by Income and Wealth Quintile, Fourteen Years Into High Medical Inflation Scenario

Income	Wealth Quintile					
Quintile	Lowest	Second	Third	Fourth	Highest	All
Lowest	-0.21%	-0.19%	-0.20%	-0.19%	-0.50%	-0.35%
Second	-0.22%	-0.15%	-0.21%	-0.31%	-0.39%	-0.30%
Third	-0.52%	-0.37%	-0.51%	-0.48%	-0.53%	-0.49%
Fourth	-0.48%	-0.51%	-0.51%	-0.59%	-0.58%	-0.55%
Highest	-0.50%	-0.54%	-0.72%	-0.76%	-0.68%	-0.66%

Table 4: Difference in Out-of-Pocket Medical Expenses Relative to Baseline by Income and Wealth Quintile, Fourteen Years Into High Medical Inflation Scenario

Income	Wealth Quintile					
Quintile	Lowest	Second	Third	Fourth	Highest	All
Lowest	7.97%	8.42%	9.08%	9.95%	10.58%	9.70%
Second	9.71%	9.86%	9.98%	10.28%	10.52%	10.16%
Third	9.61%	10.53%	10.38%	10.43%	9.89%	10.14%
Fourth	10.07%	10.04%	10.28%	10.31%	9.84%	10.08%
Highest	9.33%	9.15%	8.69%	8.38%	6.62%	8.03%

Table 5: Difference in Total Quantity of Medical Care to Baseline by Income and Wealth Quintile, Fourteen Years Into High Medical Inflation Scenario

Income	Wealth Quintile					
Quintile	Lowest	Second	Third	Fourth	Highest	All
Lowest	-4.26%	-4.28%	-4.49%	-5.17%	-5.75%	-5.04%
Second	-5.27%	-5.27%	-5.40%	-5.60%	-6.36%	-5.69%
Third	-5.81%	-6.00%	-6.25%	-6.47%	-7.13%	-6.49%
Fourth	-6.59%	-6.67%	-6.78%	-6.89%	-7.35%	-6.95%
Highest	-7.69%	-7.93%	-8.34%	-8.63%	-10.13%	-8.89%

Table 6: Difference in Health Investment Relative to Baseline by Income and Wealth Quintile, Fourteen Years Into High Medical Inflation Scenario

Income	Wealth Quintile					
Quintile	Lowest	Second	Third	Fourth	Highest	All
Lowest	-20.05%	-25.85%	-24.11%	-21.17%	-17.81%	-17.97%
Second	-18.58%	-17.74%	-17.25%	-16.70%	-16.36%	-16.55%
Third	-16.62%	-15.82%	-16.38%	-16.28%	-17.17%	-16.91%
Fourth	-16.10%	-16.29%	-16.14%	-16.59%	-16.78%	-16.58%
Highest	-16.09%	-16.32%	-16.81%	-17.08%	-16.91%	-16.84%

Table 7: High Medical Inflation Scenario Equivalent Variation (as Percentage of Initial Wealth) by Income and Wealth Quintile

Income	Wealth Quintile					
Quintile	Lowest	Second	Third	Fourth	Highest	All
Lowest	-22.12%	-40.43%	-7.93%	-2.38%	-1.52%	-1.85%
Second	-2.35%	-9.72%	-5.90%	-3.14%	-1.45%	-2.25%
Third	-27.97%	-4.97%	-3.23%	-2.19%	-1.64%	-2.25%
Fourth	-4.17%	-3.91%	-3.39%	-2.41%	-1.82%	-2.39%
Highest	-9.26%	-3.36%	-3.54%	-2.54%	-1.28%	-2.25%

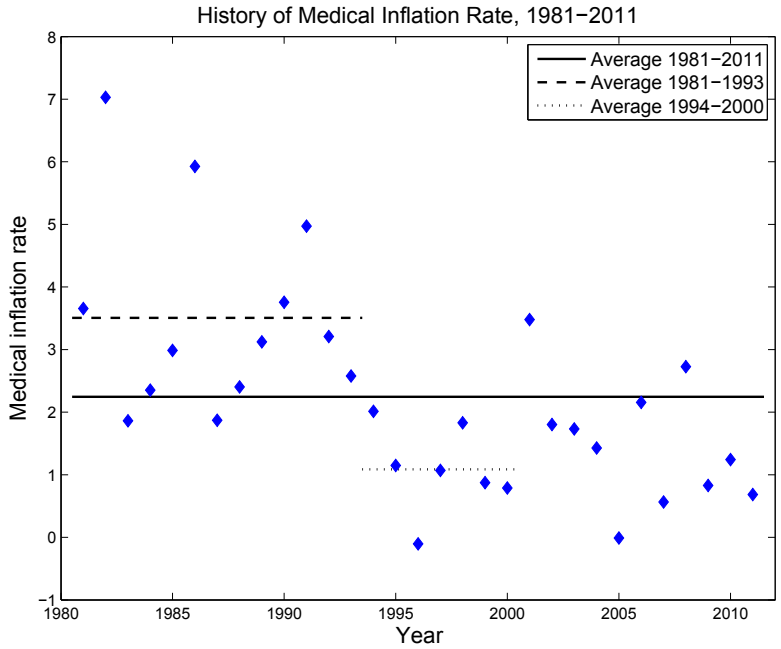

Figure 1: Medical inflation over a 30 year period. Relative price of care is calculated as the medical component of the CPI divided by the non-medical component. Inflation values calculated as annual growth rate during each year.

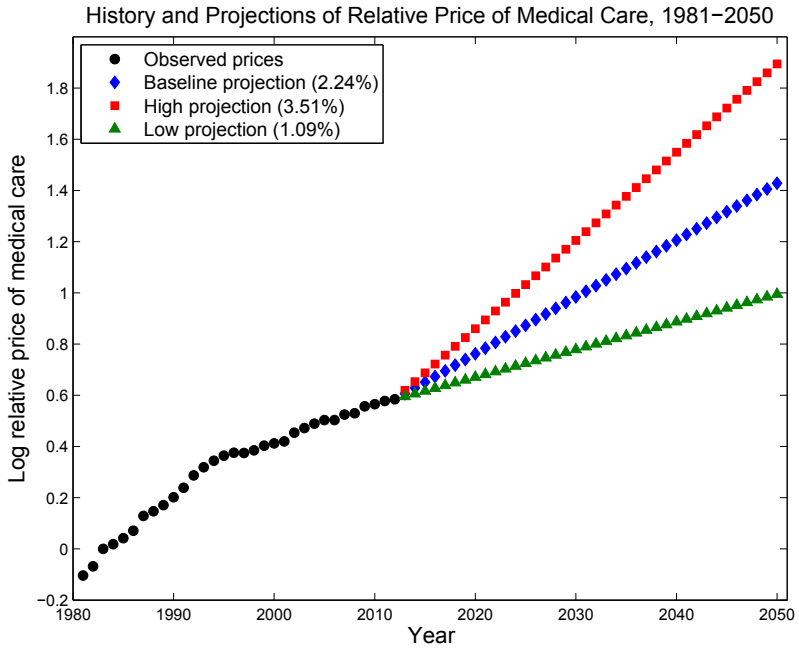

Figure 2: Observed relative price of medical care, with three projected future paths. Baseline scenario assumes long run medical inflation rate from 1980-2011; high inflation scenario uses average rate from 1980-1993; low inflation uses average rate from 1994-2000.

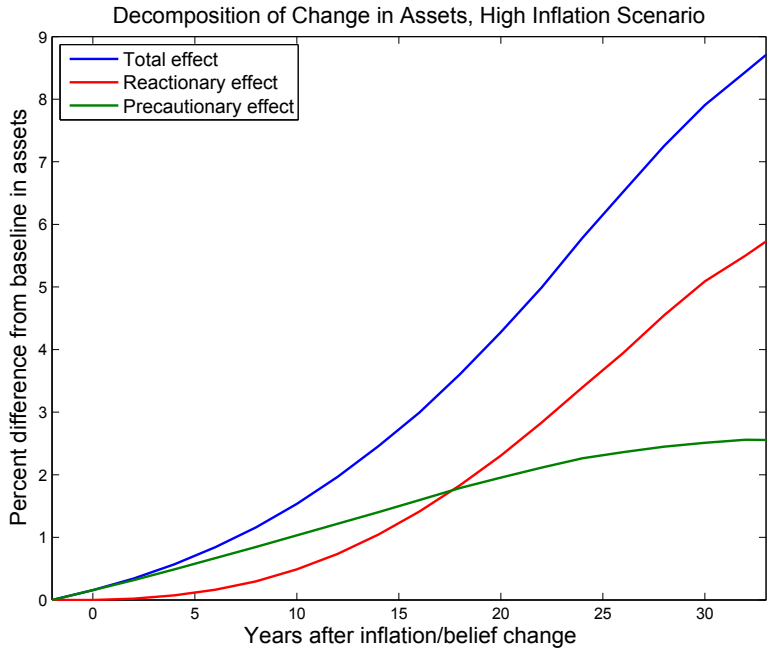

Figure 3: Percentage difference in asset holdings (relative to baseline) in three counterfactuals of the high inflation scenario. Blue path shows outcome when individuals correctly believe medical inflation is high. Red path shows outcome when inflation is high but individuals believe it is at baseline. Green path shows outcome when inflation is at baseline but individuals believe it is high.

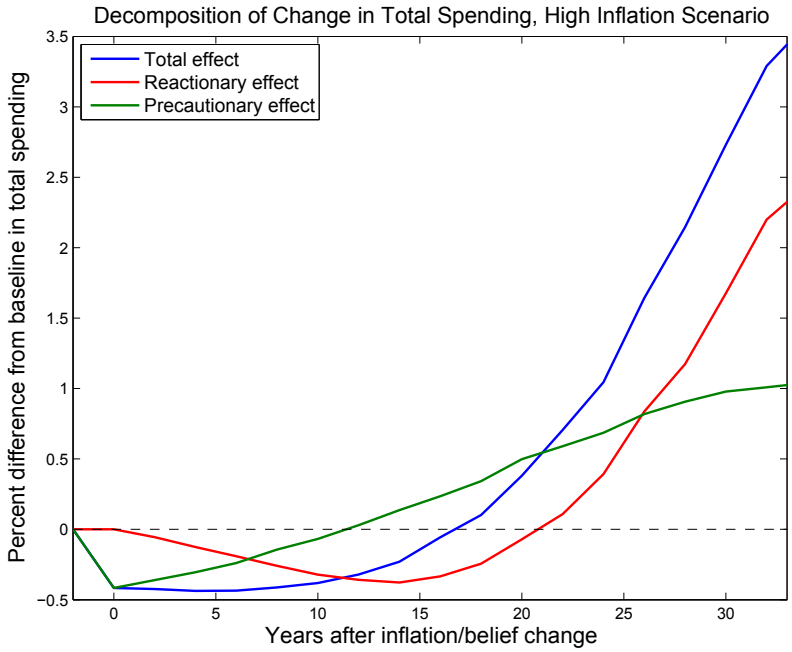

Figure 4: Percentage difference in total spending (relative to baseline projection) in three counterfactuals of the high inflation scenario.

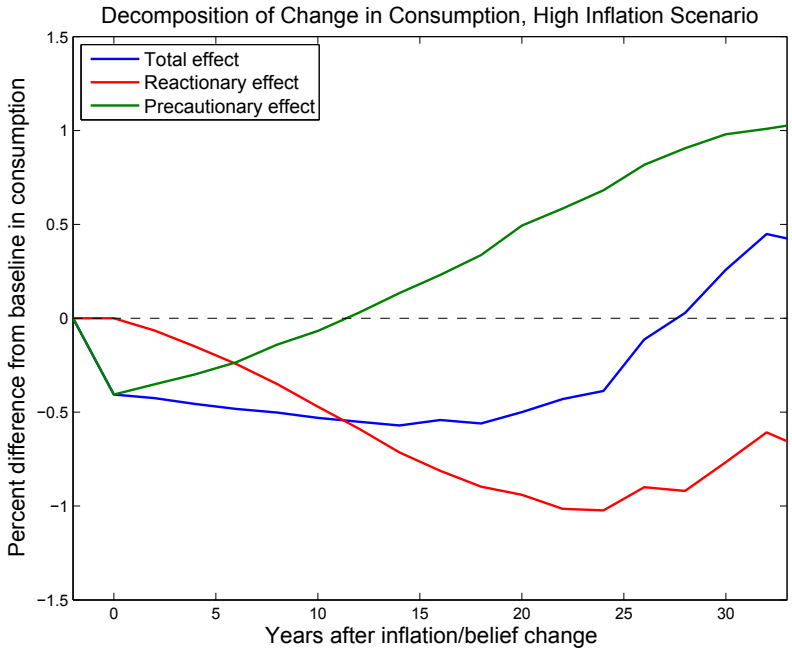

Figure 5: Percentage difference in consumption (relative to baseline projection) in three counterfactuals of the high inflation scenario.

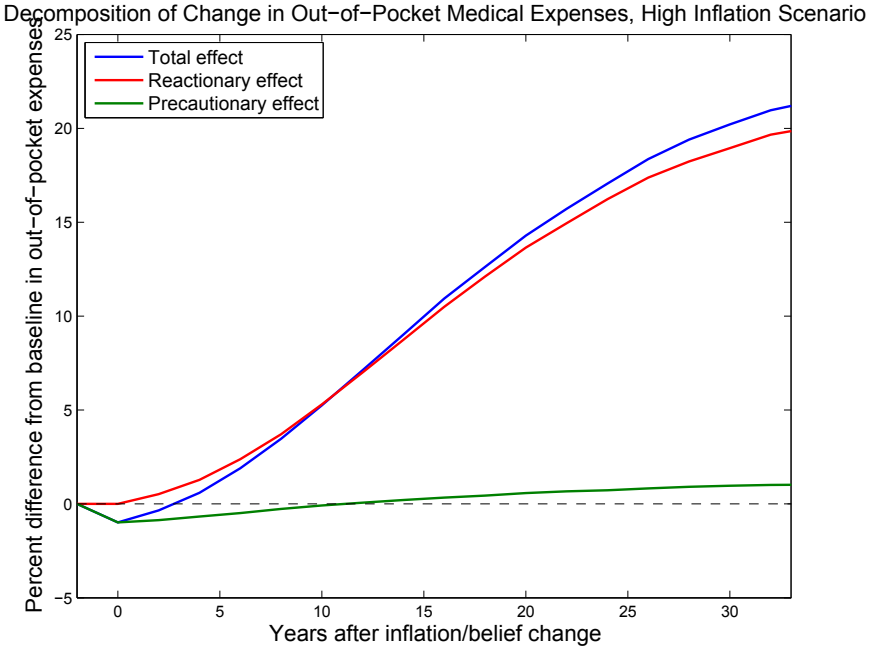

Figure 6: Percentage difference in out-of-pocket medical expenses (relative to baseline projection) in three counterfactuals of the high inflation scenario.

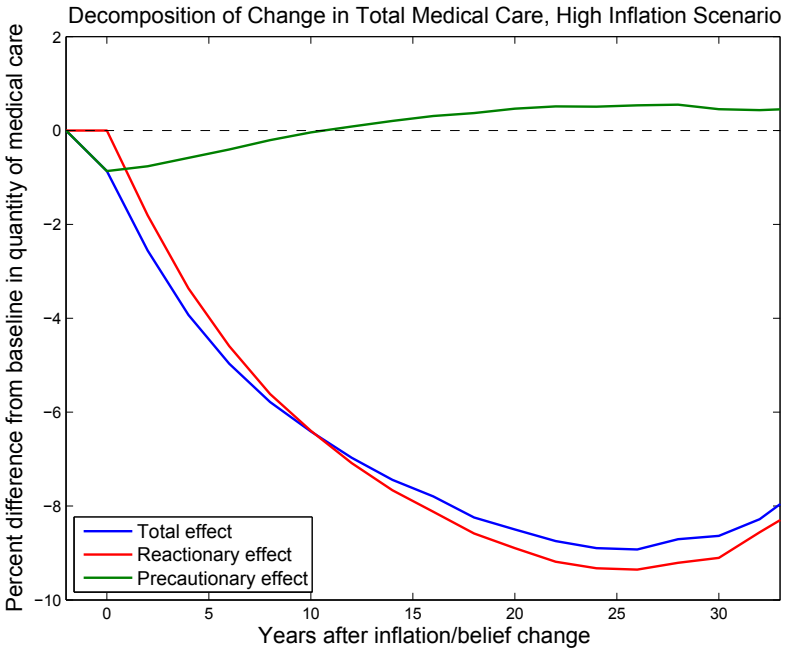

Figure 7: Percentage difference in total quantity of medical care (relative to baseline projection) in three counterfactuals of the high inflation scenario.

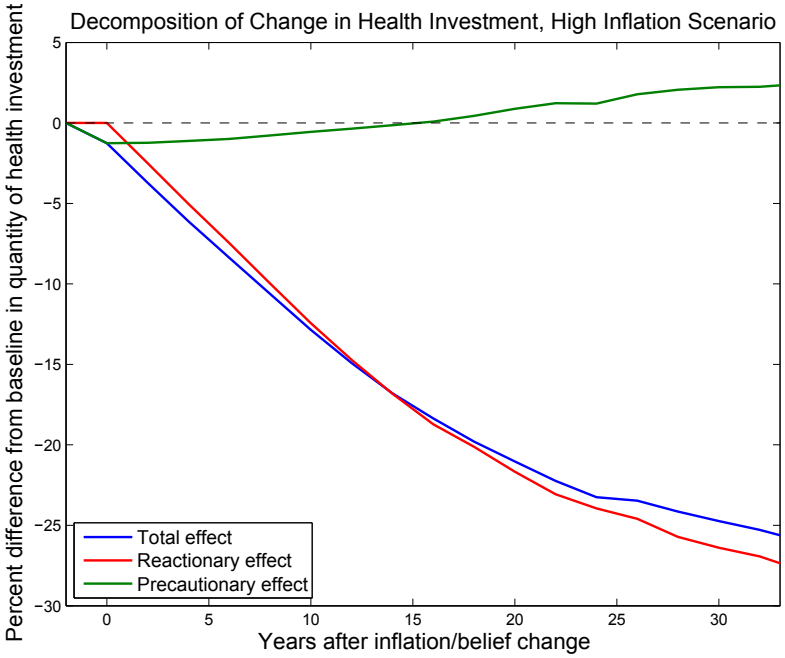

Figure 8: Percentage difference in health investment (relative to baseline projection) in three counterfactuals of the high inflation scenario.

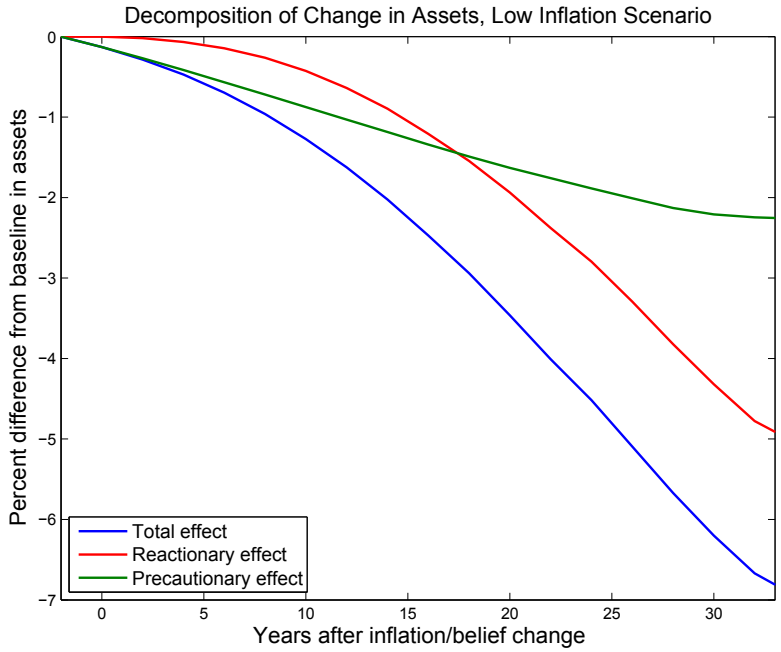

Figure 9: Percentage difference in asset holdings (relative to baseline projection) in three counterfactuals of the low inflation scenario.

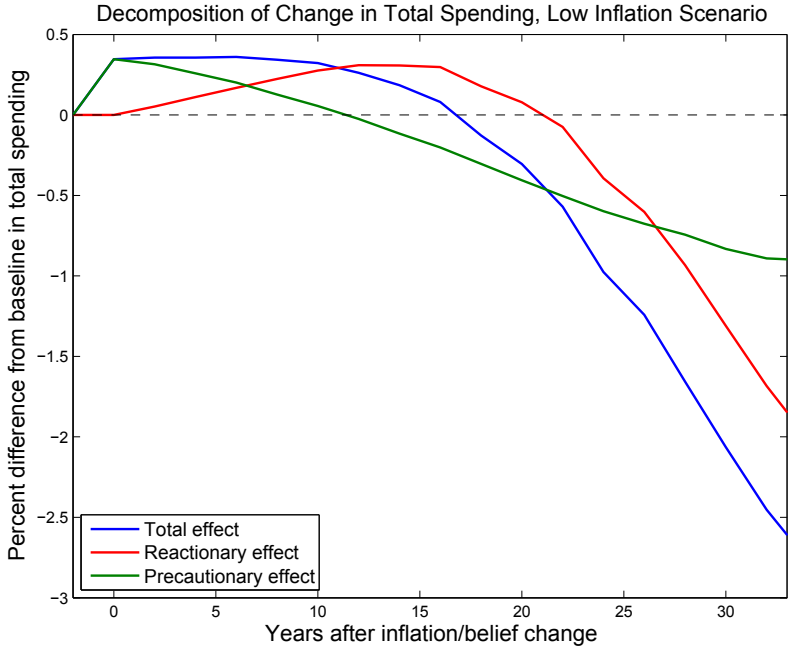

Figure 10: Percentage difference in total spending (relative to baseline projection) in three counterfactuals of the low inflation scenario.

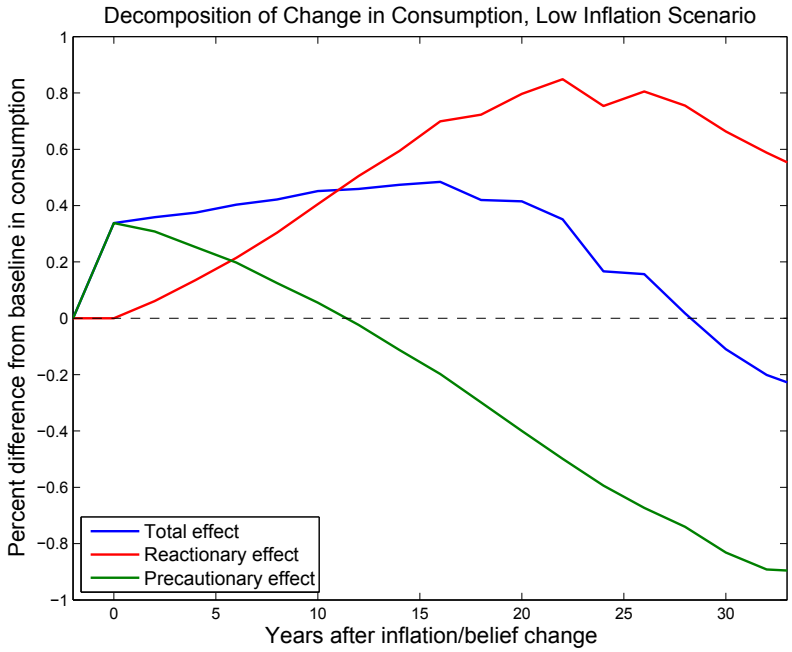

Figure 11: Percentage difference in consumption (relative to baseline projection) in three counterfactuals of the low inflation scenario.

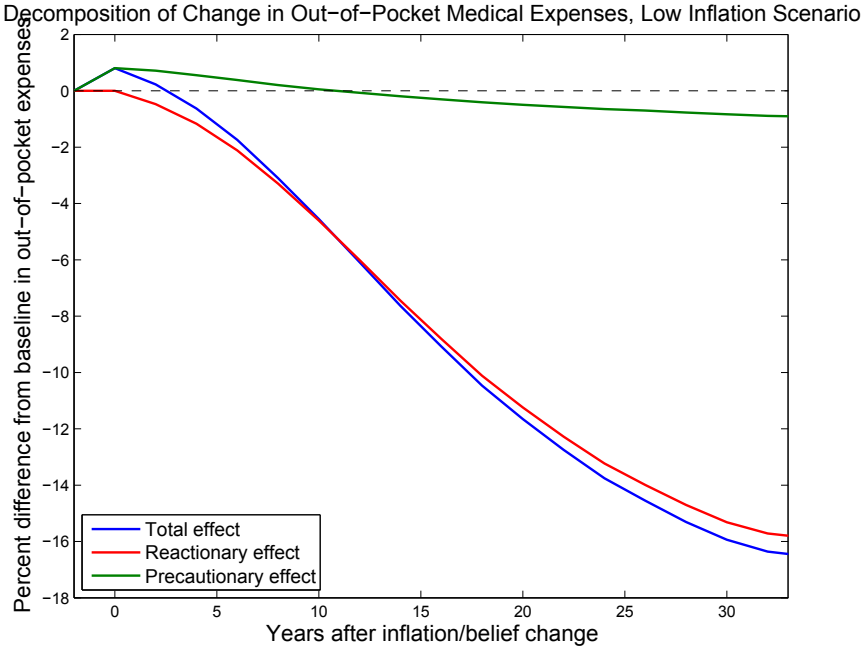

Figure 12: Percentage difference in out-of-pocket medical expenses (relative to baseline projection) in three counterfactuals of the low inflation scenario.

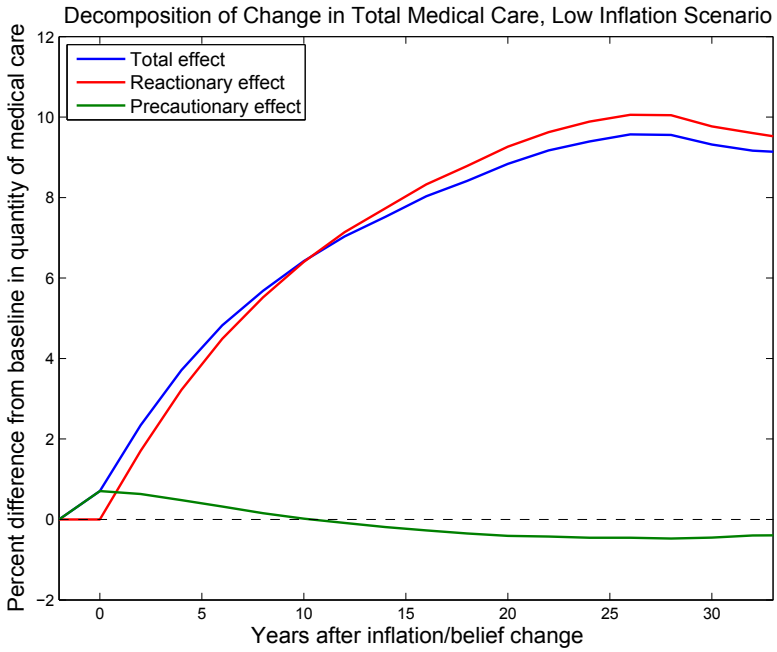

Figure 13: Percentage difference in total quantity of medical care (relative to baseline projection) in three counterfactuals of the low inflation scenario.

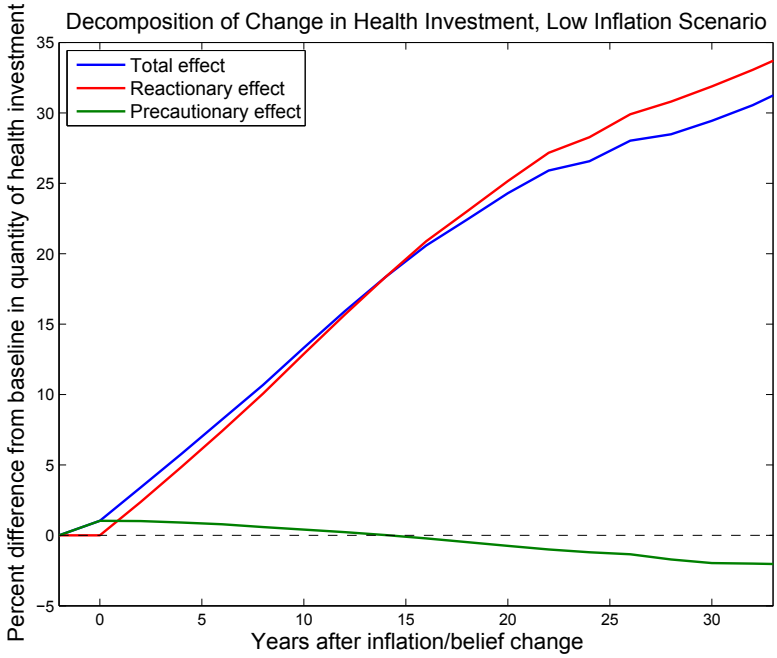

Figure 14: Percentage difference in health investment (relative to baseline projection) in three counterfactuals of the low inflation scenario.

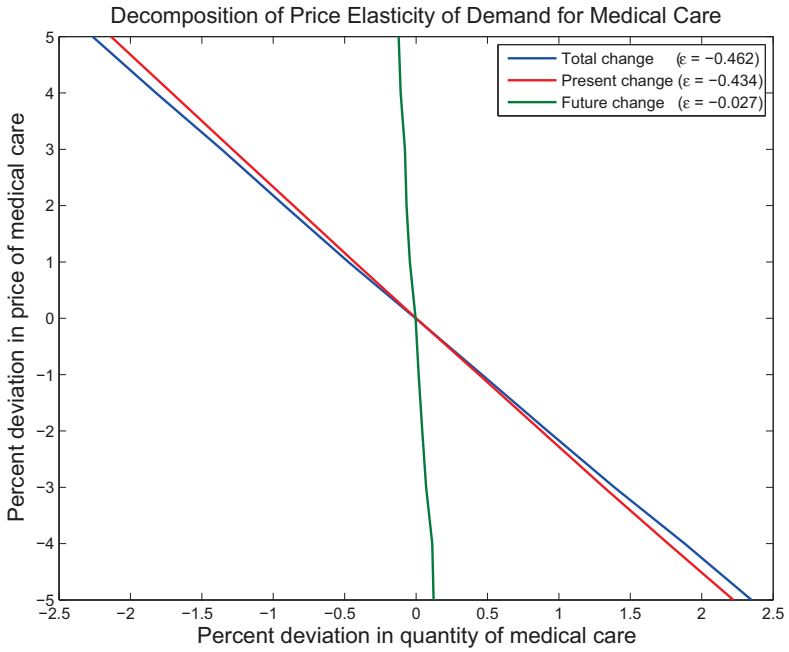

Figure 15: Static decomposition of price elasticity of demand for medical care between current and future price changes.

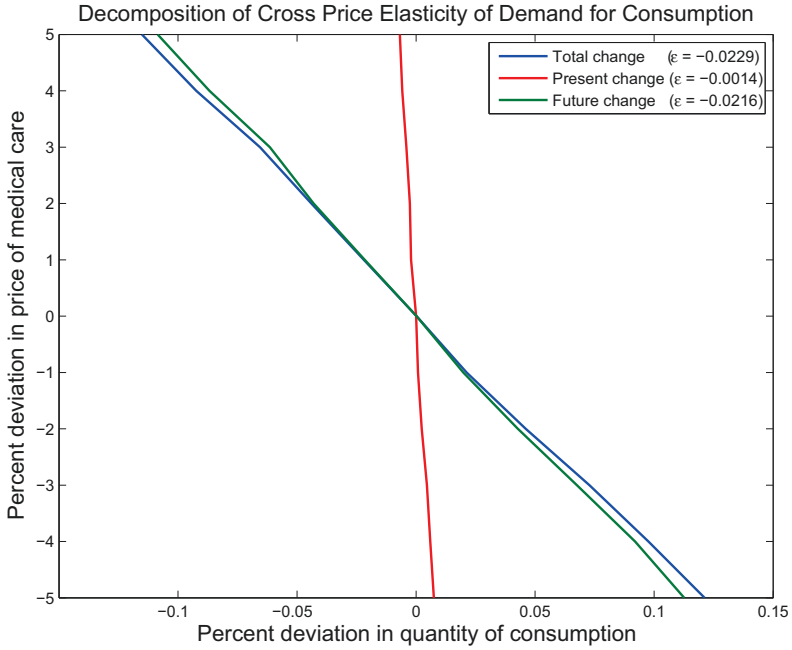

Figure 16: Static decomposition of cross price elasticity of demand for consumption between current and future price changes (for medical care).

This page intentionally left blank.

Chapter 3:
Competition Among Insurers and Consumer Welfare

Abstract

This chapter presents a model to analyze consumer welfare, price, and competition in a three-way market among patients (consumers), medical providers, and insurers. Examples are used to demonstrate that consumer welfare is not necessarily increasing in the number of insurers, but instead exhibits a U-shaped pattern. While insurers compete with each other for customers, they also act as collective bargaining agents on behalf of patients in determining the equilibrium price of health care with providers. The entry of an additional insurer thus has contradictory effects on prices and consumer welfare, reducing prices through competition but increasing them through reduced bargaining power of incumbent insurers. Moreover, the more favorable contracts allow individuals to purchase care more often, shifting out the demand curve for care and resulting in a higher equilibrium price. I characterize the equilibrium of the game under all combinations of monopoly and competition between providers and insurers. In the parameterized model, the demand effect of additional insurers combines with the cannibalization of bargaining power to overcome the competitive benefits of an entrant insurer when the number of incumbents is small, but dominance is reversed when there are more than three incumbents.

Notes: The author thanks Hülya Eraslan, who greatly aided the presentation of this chapter in terms of clarity and exposition.

1 Introduction

Recent legislative proposals to rein in the cost of medical care and medical insurance in the United States have focused on increasing competition among insurers. The Affordable Care Act of 2010 includes the creation of insurance exchanges, online marketplaces that allow consumers buying individual policies (i.e. not through their employer) to consider comparable policies from different insurers, forcing the firms to compete directly on price. Rival policies authored by Republicans would allow the interstate sale of insurance policies, increasing the number of insurers operating in local or state markets to foster more competition. To better understand the role of competition among insurers in determining the price of insurance and medical care, this chapter specifies a model of the market for health care among individuals, insurers, and medical providers. After solving for the equilibrium of the model, I analyze how the number of insurers and medical providers affects consumer welfare as well as the price of care.

1.1 Discussion

In a market for medical care where there are a relatively small number of providers, these providers have market power over individuals and are able to set their price above marginal cost. Because any given individual has no market power, he cannot negotiate for a lower price by threatening to choose a different provider. In this way, medical insurers improve consumer welfare both by smoothing utility in the face of uncertainty (from medical needs shocks) and by acting as a collective bargaining agent on behalf of individuals to set the price of care. While a single individual's threats are infinitesimal, the purchasing power of a set of individuals aggregated by an insurer provides real negotiating power. If an insurer can credibly threaten to not cover a provider's services unless that provider lowers its price of care, the equilibrium price level reached by this sort of bargaining will be significantly lower.

The rationale behind legislation to increase competition among insurers relies on the fact that when there are a larger number of firms competing for customers, they will offer insurance contracts more favorable to consumers (the "competitive effect"). However, there are also two countervailing effects from the entry of an additional insurer. First, the cannibalization of market share and thus a decrease in the threat power of both the incumbent and entrant insurers when negotiating prices with medical providers (the "bargaining effect"). Second, an outward shift in the demand curve for care due to the more favorable contracts, which increases the price of care (the "demand effect"). A larger number of competing insurers will generate more favorable insurance contracts conditional

on the price of care, but the equilibrium price will be higher. In this way, the effect on consumer welfare from the entry of an additional insurer is ambiguous.

The model presented here includes both competition among insurers and a form of negotiation between insurers and providers so that all three effects emerge. In the model, providers each choose a price at which to sell care; after observing these prices, each insurer chooses a contract to offer, modeled as a premium and a copay for each provider. Upon seeing the menu of offered contracts, individuals choose up to one contract, learn their medical need, and decide whether to buy care. Individuals' medical needs are drawn from a known distribution and are private information to that individual; an individual's need is experienced as a penalty to utility unless the individual purchases a unit of care. To elucidate the effect of bargaining power in determining equilibrium prices, two subgame perfect Nash equilibria are described. Consumers' strategies and the terms of the contracts offered by insurers (the premium and lowest copay) do not vary between the two equilibria. However, the two equilibria differ on insurers' strategies for including providers in their coverage network, with insurers strategically withholding inclusion as a form of bargaining by threat.

In the pure strategy "simple equilibrium", insurers include in their coverage network any provider offering the lowest price among rivals. Providers in the model engage in Bertrand competition on price, but face increasing marginal costs of producing care. This prevents providers from fully competing to the point where price equals marginal cost, with a symmetric equilibrium reached at a higher price level. In contrast, the mixed strategy "threatening equilibrium" has insurers randomly withhold coverage of each provider's services when all providers offer the same price, creating variance in quantity of care demanded from that provider. Because the marginal cost of care is increasing, provider profit is concave with respect to quantity of care produced. The variation in demand thus reduces expected profit and induces providers to compete to prices lower than they could when insurers employ the simple equilibrium strategy. While both strategies admit a range of equilibrium prices with identical lower bounds, the upper bound on this range is lower for the threatening equilibrium. Further, the upper bound on equilibrium prices is increasing in the number of insurers and decreasing in the number of providers.

An analysis of consumer welfare under an example parameter set demonstrates that the overall effect on welfare from additional insurers is not always monotonic, but instead can follow a U-shaped pattern. The loss of bargaining power caused by an additional insurer is large when the number of incumbents is small; likewise, the shift in total demand for care is also larger when the number of insurers is small. Thus the entry of the second or third insurer can have a net negative impact on

individuals' welfare if these effects overwhelm the benefit of a more favorable insurance contract due to increased competition among insurers. The amount of bargaining power lost decreases with each entering insurer, and the total demand function similarly converges as the number of insurers grows. This allows the benefits of more competition to dominate, increasing individuals' welfare when the number of incumbent insurers is relatively large. In contrast to the entry of an additional insurer, reducing barriers to competition (represented in the model as the variance of preference shocks over insurance contracts) among incumbent insurers has an unambiguously positive effect on welfare so long as at least two insurers compete.

The U-shaped welfare pattern is not universal, but rather depends on the parameters of the model. When the magnitude of medical needs shocks is doubled, for example, individuals' welfare monotonically increases with the number of insurers when there are a relatively large number of providers. On the other hand, when the variance of preference shocks over insurance contracts is large (so that individuals are "poor shoppers", unable to accurately compare contracts), individuals' welfare monotonically decreases in the number of insurers in many cases. The baseline parameters have not been selected to match real world data, so the patterns seen in the analysis are meant to demonstrate the complexity of the interactions in the three-way market rather than to predict the outcome of specific policies.

Each of the three effects of insurer entry is directly captured in the model by a specific component. The three effects can be isolated by solving for the equilibrium of similar games in which particular components are removed or held constant, allowing the total effect of insurer entry to be decomposed. For example, the pattern of welfare with respect to the number of insurers can be considered when the bargaining effect is "turned off" by eliminating insurers' threat power entirely (or by allowing insurers to collude when bargaining, maintaining full threat power even with multiple insurers). This decomposition reveals that the possible net loss of consumer welfare from the entry of an additional insurer does not require both the demand effect and bargaining effect: even if either is turned off, the U-shaped welfare pattern persists. When both effects are turned off, however, individuals' welfare is strictly increasing in the number of insurers as well as medical providers, as only the competitive effect is present. This scenario would occur if only one consumer were offered the opportunity to purchase insurance by several competing insurers; no shift in the demand curve would occur, nor would insurers have market power to negotiate with medical providers. This decomposition shows that individually rational choices to purchase insurance can aggregate to be collectively harmful through changes to the price of care induced by moral hazard.

To emphasize how strong a role moral hazard plays in increasing the equilibrium price of care, an alternative specification considers a world in which medical needs are observable and contractible. In the baseline model, insurers know that reducing individuals' out-of-pocket cost of care will cause them to purchase care more often and for less severe medical needs than when they pay full price. A complete insurance contract (one with no copay) is thus impossible, as individuals would obtain care all of the time, even for the smallest medical conditions.[1] In the alternative model, insurance contracts are complete in equilibrium, but specify a minimum level of medical need for the care to be reimbursed. Repeating the welfare analysis reveals that containing moral hazard in this way severely limits the outward shift in demand from additional insurers, while the competitive effect is enough to overcome the loss of bargaining power. Additional insurers strictly improve welfare (except in one corner case), and prices are significantly lower than in the baseline model. In combination with insurance contracts being complete, with no out-of-pocket cost, welfare is almost universally higher when medical needs are observable than when the demand and bargaining effects are turned off in the baseline model.

1.2 Related Literature

The notion that insurers can use their market power to negotiate with hospitals and physicians for lower service prices is not novel, and goes back at least as far as the 1980s. Moreover, moral hazard in health insurance markets has been well understood for decades. While there have been empirical studies of insurer competition (or insurer market concentration) on negotiated service prices and premiums offered to customers, to date no paper has combined all of these aspects into a single theoretical model. In particular, previous work has been silent on the generosity of insurance contracts (through lower copayments) with respect to competition among insurers and providers, and its effect on consumer welfare. This section provides an overview of the empirical literature on insurer competition and bargaining with providers and a brief discussion of the closely related literature on competition among hospitals. It further describes previous theoretical work on insurance network formation games and upstream-downstream bargaining games.

A common theme in work on setting hospital service prices is that an insurer must be able to channel its customers to particular hospitals in order to have bargaining power and reduce price. Staten, Dunkelberg, and Umbeck (1987) examine whether Blue Cross can force hospitals to offer

[1] In the stylized model, there is no time cost of medical care, only monetary cost. In reality, free health care would not actually result in 100% utilization.

discounted prices via their large market share and find no relationship between the insurer's market share and the discount extracted on a hospital-by-hospital basis. They argue that BC is unable to direct its patients to specific hospitals as a punishment/reward mechanism in negotiations. With the rising popularity of managed care in the late 80s and early 90s, however, HMOs and PPOs were able to channel their patients to particular hospitals and physicians through selective contracting. Melnick, Zwanziger, Bamezai, and Pattison (1992) present evidence to show that negotiated prices are increasing in the importance of a hospital to patient preferences and decreasing in hospital competition in a local area. With more complete national data, Bamezai, Zwanziger, Melnick, and Mann (1999) also find evidence supporting the importance of patient channeling; Wu (2009) argues that managed care organizations engage in "partial channeling" by channeling patients within a broad network.

Overall, the empirical literature is fairly mixed on the issue of whether increases in competition among insurers leads to lower hospital service prices or premiums. Melnick, Shen, and Wu (2011) find a negative association between negotiated hospital prices and concentration in insurer markets, and a positive association with hospital concentration. Bates and Santerre (2008) consider whether insurers use their market power to bust the near monopolies in concentrated provider markets or instead act as monopsonist intermediaries. They find that there is no evidence that insurer market concentration is used against consumers as monopsony power, and some evidence that health services output is actually increasing in insurer concentration. In contrast, Dafny (2010) presents evidence that insurers do not act as benevolent agents for consumers, passing all service price reductions on to customers in the form of reduced premiums, but instead engage in direct price discrimination that is only possible in imperfect markets. Moreover, Dafny, Duggan, and Ramanarayanan (2009) examine the effects of the merger between two very large national insurers on premium growth rates in local markets, finding that exogenous increases in concentration are associated with an increase in premium growth. Very recent work in Ho and Lee (2013) develops a theoretical bargaining model and shows that increased competition among insurers reduces negotiated hospital prices, except for those that are most attractive to consumers. A working paper by Trish and Herring (2014) separately estimates the effect of insurer market concentration on premiums for "fully insured" employer-sponsored policies vs "administrative services only" contracts. They find that when insurers only manage health benefits (and thus do not negotiate prices with hospitals), concentration is positively associated with premiums, and that this relationship reverses in the presence of negotiation.

The literature on competition among medical providers universally finds that greater concentra-

tion leads to higher prices, as is typically expected. Capps and Dranove (2004) find that mergers among hospitals significantly raise negotiated prices in most cases. Less strongly, Haas-Wilson and Garmon (2011) find anticompetitive effects from one of the two hospital mergers that they study in close detail, with large price increases post-merger. Gaynor and Vogt (2003) estimate a structural model of for-profit and non-profit hospitals' pricing decisions and then simulate the effects of mergers between hospitals of the same type, finding no difference in post-merger price increases between the two types. Taking a more theoretical approach, Capps, Dranove, and Satterthwaite (2003) model consumer preferences for hospitals as a discrete choice conditional on diagnosis; an insurer thus aggregates hospitals into a coverage network that acts as an option contract. Ho (2009) presents related evidence that individuals strongly weigh an insurer's provider network when selecting a plan, and selective contracting leads to consumer welfare loss of about $1B per year.

This paper's model includes a form of bargaining between insurers and providers in which the number of firms is fixed and a complete network is formed by all insurers in equilibrium, but the theoretical literature presents some alternative possibilities. Inderst and Wey (2003) analyze a model with two "upstream" suppliers selling to two "downstream" retailers, deriving conditions under which each side of the market would prefer to merge into a single firm. Gal-Or (1997) presents a related "two on two" model of insurer-provider bargaining, focusing on when there will be an exclusionary equilibrium in which neither insurer contracts with a particular hospital. She finds that the exclusionary equilibrium is the only one possible when consumer preferences for additional choice (due to provider differentiation) is sufficiently small relative to customer attachment to the insurer, while agreements will be reached by all insurers and hospitals when the reverse is true.

The particular form of bargaining employed in this chapter's model, in which an insurer uses the threat of complete and random patient channeling to introduce variance in quantity demanded from each provider, is unique to the literature (though weaker forms of it are hinted at in descriptive studies). However, there is no standard bargaining model employed universally in empirical studies that motivate their estimations with theoretical models. Nash bargaining seems to be a default choice, as in Ho and Lee (2013) and other papers. While the existence of equilibria can be demonstrated in these models, uniqueness is a persistent issue. Moreover, equilibria can often be characterized, but fully solving for an exact solution is difficult in a simultaneous bilateral negotiation model, as would be the case with N insurers and M medical providers. Alternative bargaining models include the take-it-or-leave-it approach in Ho (2009), with hospitals as the side making the single offer, or a sort of "reduced form Nash" approach in Capps, Dranove, and Satterthwaite (2003) in which

the gains from trade are split between the insurer and provider in a fixed proportion. The unique form of bargaining presented here was selected specifically for its tractability and clean, closed-form solution, while generating the key results common in empirical studies.

The work presented in this chapter contributes to the existing literature in several ways. As noted above, it employs a form of bargaining not previously modeled, an extreme form of the patient channeling strategies described in empirical studies of insurer concentration. More importantly, it is the first theoretical model that simultaneously captures several effects that have been modeled or discussed separately in previous work: competition among insurers potentially reducing premiums, insurer concentration bolstering bargaining power, and moral hazard from medical insurance. Empirical work concerned with insurer-provider bargaining tends to focus on negotiated prices while ignoring the premiums passed on to consumers, while studies of insurer competition make the opposite choice. Both strands of literature nearly universally ignore the role of moral hazard– well documented yet elsewhere, in studies of the elasticity of demand for medical care. To that end, the model presented in Section 2 is also somewhat unique in explicitly accounting for the generosity of insurance contracts (through the copay) and solving for the equilibrium terms of the insurance contract. While no individual component of the model is fully novel, this chapter synthesizes several previously identified phenomena into a single tractable model that allows an analysis of their interaction and effects on consumer welfare itself– rather than indirect proxies like premiums or negotiated prices– through several channels.

The structure of this paper is as follows: Section 2 presents the model, Section 3 solves for the subgame perfect Nash equilibrium of the market, Section 4 analyzes consumer welfare in the model, and Section 5 concludes.

2 Model

This section describes a one-shot game representing a market for medical care and insurance. The players consist of a unit mass of individuals indexed by i, a finite number of insurers indexed by $j = 1, 2, ..., N$ and a finite number of medical providers indexed by $k = 1, 2, ..., M$. Individuals use consumption and medical care to maximize their expected utility, insurers sell insurance contracts to individuals to maximize their expected profit, and medical providers produce medical care to sell to individuals (potentially intermediated by insurers) to also maximize expected profit.

2.1 Timing

The order of events during the game is as follows:

1. Each of the M medical providers simultaneously chooses a price $p_k \in \mathbb{R}_+$ at which to sell medical care.

2. Each of the N insurers learns the vector of prices $\vec{p} = (p_1, \cdots, p_M)$, then they each simultaneously choose an insurance contract $\hat{\chi}_j \in \mathbb{R}_+^{M+1}$ to offer, representing the copays for each provider and the premium.

3. Individuals see the menu of offered insurance contracts $\hat{X} = (\hat{\chi}_0, \hat{\chi}_1, \cdots, \hat{\chi}_N)$ and choose one to purchase (possibly a null contract), paying the premium for that contract.

4. Individuals learn their level of medical need, decide which provider to purchase care from (possibly none) and how much to consume.

5. Each medical provider produces care to meet individuals' purchasing decisions, incurring production costs and gaining revenue from sales at their chosen price. Care is paid for by insurers and individuals at the contracted cost sharing.

2.2 Individuals

Individuals have a common utility function over consumption $u(x)$, with $u'(x) > 0$ and $u''(x) \leq 0$ for all $x > 0$; as a technical condition, $\lim_{x \to 0^-} u'(x) = \infty$, which is satisfied by constant relative risk aversion (CRRA) utility with coefficient of risk aversion ρ. Each individual i has a medical need or pain shock η_i, which is continuously distributed on $\mathbb{R}_+$ according to CDF $F(\eta)$ with associated PDF $f(\eta)$. Pain is experienced as a penalty to utility unless the individual obtains a unit of medical care $m \in \{0, 1\}$, negating the pain. For each insurance contract (including the null contract described below) in a menu $\hat{X} = (\hat{\chi}_0, \hat{\chi}_1, ..., \hat{\chi}_N)$, individual i has an associated personal preference shock ϵ_{ij} drawn from a logistic distribution with standard deviation σ; he receives this shock as a bonus to utility if he purchases that contract. An individual's total utility is thus given by:

$$U(x_i, \eta_i, m_i, \epsilon_{ij}) = u(x_i) - \eta_i(1 - m_i) + \epsilon_{ij}. \tag{1}$$

An individual does not learn his η_i until after he chooses an insurance contract to purchase, whereupon it is private information unobservable to providers or insurers. The preference shocks for each

insurer are known before the individual selects an insurance contract. Let $Dik_k = 1$ when the individual purchases care from provider k and $D_{ik} = 0$ if he does not ($D_{i0} = 1$ when the individual does not buy care at all).

Each individual has access to $y > 0$ in income or financial resources. Consumption can be purchased at a unit cost of 1, while the out-of-pocket cost of medical care depends on the individual's insurance contract. If the individual has insurance contract $\hat{\chi} = (z, \vec{c})$, then the unit of medical care costs c_k if he buys care from medical provider k. To have contract $\hat{\chi}$, the individual paid insurance premium z, reducing his resources with which to purchase consumption whether or not he buys care. The null contract $\hat{\chi}_0 = (0, \vec{p})$ is always available, where $\vec{p}$ is the vector of prices charged by the medical providers. Conditional on out-of-pocket cost, an individual has no preferences among medical providers.

When choosing an insurance contract, the individual's problem is to:

$$\max_{j \in \mathbb{N}_N} \mathrm{E}[U(x_i, \eta_i, m_i, \epsilon_{ij})]. \tag{2}$$

The expectation is taken over the distribution of medical needs that the individual could experience. After the contract $\hat{\chi}_j$ is selected (and trivially defining $c_{j0} = 0$), the level of need η_i becomes known and the individual's problem is:

$$\max_{x_i, k \in \mathbb{N}_M} U(x_i, \eta_i, m_i, \epsilon_{ij}) \text{ s.t. } x_i + c_{jk} + z_j \leq y,\ m_i = \mathbf{1}(k > 0). \tag{3}$$

2.3 Insurers

Each insurer j simultaneously chooses a single insurance contract $\hat{\chi}_j = (z_j, \vec{c}_j)$ to offer in order to maximize profit. When offering contract $\hat{\chi}_j$, the expected profit from each individual who buys the contract is:

$$r(\hat{\chi}_j) = z_j - \sum_{k=1}^{M} \mathrm{Prob}(D_{ik} = 1 | \hat{\chi}_j)(p_k - c_{jk}). \tag{4}$$

Defining $q(\hat{\chi}_j | \hat{X}_{-j})$ as the proportion of individuals who select $\hat{\chi}_j$ when rival firms (including the "null firm") offer $\hat{X}_{-j}$, then an insurer's expected profit is:

$$\pi(\hat{\chi}_j | \hat{X}_{-j}) = r(\hat{\chi}_j) \cdot q(\hat{\chi}_j | \hat{X}_{-j}). \tag{5}$$

2.4 Medical Providers

Medical providers have a common production cost function $\kappa(D_k)$ that is increasing and convex in the quantity demanded from that provider D_k. To simplify analysis, assume there are no fixed costs, $\kappa(0) = 0$, and that the first units of care are very easy to produce, $\kappa'(0) = 0$. A provider must sell to all customers who wish to purchase care at the price chosen by that provider. When provider k sells care at price p_k and D_k individuals want to buy care, it receives expected profit of:

$$\hat{\pi}_k = p_k D_k - \kappa(D_k). \tag{6}$$

2.5 Equilibrium Defined

A strategy for a medical provider is a choice of $p_k \in \mathbb{R}_+$. A strategy for an insurer is a function $\phi : \mathbb{R}_+^M \to \mathbb{R}_+^{M+1}$ that maps vectors of prices offered by providers into a choice of insurance contract to be offered. A strategy for an individual is a function $\psi^c : \mathbb{R}_+^{N+M+NM} \times \mathbb{R}^{N+1} \to \{0, 1, \cdots, N\}$ that selects an insurance contract from a menu (taking account of his preference shocks) and a function $\psi^a : \mathbb{R}_+^{M+2} \to \mathbb{R}_+ \times \{0, 1, \cdots, M\}$ that maps the insurance contract and medical needs into a choice of consumption and from which provider to purchase care (if any).

We seek a subgame perfect Nash equilibrium for the insurance market game described above. As will be shown in Sections 3.2 and 3.3, the only SPNE are symmetric and have a complete network. In this case, symmetry means that all players of the same class (individuals, insurers, and providers) choose the same strategy. An equilibrium with a "complete network" is one where equilibrium behavior results in an individual having strictly positive ex ante probability of purchasing care from any provider via any insurer, before medical needs or personal preferences are realized. In an equilibrium, individuals maximize their utility by selecting an insurance contract from the offered menu, and then optimally choose whether to purchase care once their medical need is known; each insurer offers a contract that maximizes its expected profit when individuals obey their equilibrium strategy, holding fixed the prices set by each provider and the contracts offered by rival insurers; and each provider chooses a price that maximizes their expected profit when insurers and individuals obey their equilibrium strategies, holding fixed the prices set by rival providers.

2.6 Discussion

Individuals' preference shocks over insurance contracts have several possible real world analogues. First, the shocks could represent imperfect information or difficulty understanding contracts. While the model presents a relatively simple environment, it acts as a stand-in for the much more complex insurance contracts that real consumers face. Inability to precisely compare two or more insurance contracts would lead to "choice error", allowing worse contracts to be purchased by some individuals. Alternatively, real individuals might face search costs to learn about insurance contracts, leading some of them to settle for "good enough" contracts. Both of these explanations are employed by Maestas, Shroeder, and Goldman (2009) in their analysis of price dispersion in heavily standardized Medigap insurance policies. Similarly, Frank and Lamiraud (2009) present evidence that consumers make less optimal decisions over health insurance contracts as the number of options becomes very large, and are likewise less willing to switch policies even when considerable savings are possible. Related, the fact that many individuals are offered medical insurance through their employer acts as a sort of preference shock, making certain insurers more accessible to an individual. Under any of these interpretations, we can consider how improving individuals' information or access to competing plans affects equilibrium outcomes; that is, the analysis will consider how the variance of preference shocks affects consumer welfare and the price of medical care.

One could argue that, in reality, individuals have heterogeneous preferences over medical providers as well as insurers. Any given household might live significantly closer to one hospital or another, reducing travel costs, or a patient might have an existing relationship with a doctor at a particular provider. For example, the theoretical model in Gal-Or (1997) is effectively a two-dimensional Hotelling spatial differentiation setting, with the "length" of each dimension representing the degree of attachment to insurers or providers. These preferences undoubtedly exist, and might even have larger variance than the heterogeneity of preferences with respect to insurers. Their inclusion in a theoretical model, however, greatly complicates the analysis that characterizes equilibrium and how it changes with the number of insurers and medical providers– indeed, Gal Or finds no pure strategy bargaining equilibrium in many circumstances. As this model is meant to provide a tractable framework with a particular focus on the several effects of insurer entry, this type of heterogeneity is omitted. The most important property is that both insurers and providers compete with each other in an imperfect market that allows positive profits for all in equilibrium.

3 Model Equilibrium

The model will be solved backwards, starting from individuals' choice of consumption and from which provider to purchase care (if any), and then individuals' choice of insurance contract from a fixed menu. Taking individuals' equilibrium strategy as fixed, I then find insurers' equilibrium choice of "reduced form" insurance contract to offer conditional on providers' prices (which does not depend on their strategy for their provider network). Finally, I solve for the equilibrium price that providers will offer under the "simple" and "threatening" strategies for insurers to include providers in their network.

3.1 Individuals' Equilibrium Strategy

Equilibrium behavior for the individual is straightforward to find. Because any given individual has no strategic threat value, each individual simply tries to maximize his own utility conditional on the menu of contracts offered.

Claim 1 *After an insurance contract has been selected, an individual purchases care if and only if it would result in higher utility given his realized pain, consuming all remaining resources. If care is purchased, it is from a provider with the lowest copay in the individual's contract.*

The individual will consume any resources not spent on the premium or copay because $u'(x) > 0$. Moreover, an optimizing individual who purchases care will do so from the provider with the lowest out-of-pocket price, as doing otherwise would reduce his payoff from consumption. For any insurance contract, only the lowest copay value is relevant, so we can write the "reduced form contract" of $\hat{\chi}$ as $\chi = (z, \min(\vec{c})) = (z, c)$. An individual can randomize among any providers with the lowest copay.

Claim 2 *Individuals holding reduced form contract χ purchase care with probability*
$\text{Prob}(m_i = 1|\chi) = 1 - F(u(y - z) \ - u(y - z - c))$.

Conditional on having contract $\chi = (z, c)$, he will choose the higher utility between purchasing and not purchasing care, receiving a payoff of:

$$\max\{u(y - z) \ - \eta, u(y - z - c)\} = \max\{u_0 \ - \eta, u_1\}. \tag{7}$$

Defining $\Delta u = u_0 - u_1$, an individual buys care when $\Delta u > \eta$, which occurs a fraction $1 - F(\Delta u)$ of the time. That is, $\text{Prob}(m_i = 1|\hat{\chi}) = 1 - F(\Delta u)$, which will be relevant for an insurer's decision problem. Temporarily omitting the personal preference shock ϵ_{ij}, if an individual purchases contract χ, his expected utility is:

$$\sigma\overline{u}(\chi) = F(\Delta u)u_0 + (1 - F(\Delta u))u_1 - \int_0^{\Delta u} \eta f(\eta)d\eta = u_1 + F(\Delta u)\Delta u - \int_0^{\Delta u} \eta f(\eta)d\eta. \quad (8)$$

That is, he receives u_1 the portion of time that he purchases care, u_0 the portion of time that he does not purchase care, and experiences pain when he does not purchase care.[2]

Claim 3 *An individual selects the contract that offers him the highest expected utility before his medical needs are known.*

We can now formally define the two functions that compose individuals' equilibrium strategy. To select an insurance contract, the individual maximizes his total utility among the menu of choices $\hat{X} = (\chi_0, \chi_1, \cdots, \chi_N)$, taking account of his preference shocks among these contracts:

$$\psi^c(\hat{X}, \epsilon_i) = \arg\max_j \overline{u}(\chi_j) + \epsilon_{ij}. \quad (9)$$

Defining the operator $\text{Rand}(\cdot)$ as a random selection from the given set, the individual's optimal consumption and medical care decision rule is:

$$\psi^a(\hat{\chi}, \eta_i) = \begin{cases} (y - z, 0) & \text{if } \eta_i < \Delta u \\ (y - z - c, \text{Rand}(\underline{K})) & \text{if } \eta_i \geq \Delta u \end{cases}, \quad \underline{K} = \{\arg\min_k c_k\}. \quad (10)$$

3.2 Insurers' Equilibrium Strategy

Turning now to the equilibrium behavior of insurers, this section first characterizes an insurer's best response to his rivals' choices, then demonstrates the existence of the (very likely unique) symmetric equilibrium contract for each price offered by insurers, and finally formalizes this equilibrium to match the definition provided in Section 2.5. Individuals are assumed to follow the strategy above, and the vector of prices is treated as fixed at some $\vec{p}$ already chosen by providers.

[2]The (inverse) factor of σ is included here so that the insurance contract choice probabilities follow the typical formula. Inversely scaling the expected utilities is equivalent to scaling the magnitude of preference shocks.

3.2.1 Insurers' Profit and Best Response

Claim 4 *Insurer profit as a function of its own contract χ_j is an analytic function of the premium z_j and the smallest copay offered $\min(\vec{c}_j)$, and only depends on rival insurers' contracts through a summary statistic: the sum of exponentiated expected utilities of those contracts.*

With individuals' equilibrium behavior known, we can refine the functional definitions laid out above. Because individuals will buy care based on the lowest copay rate, it can never be optimal to offer the lowest copay rate for providers charging different prices; that strategy is always dominated by one that raises the copay rate on a provider not offering the lowest price but getting the lowest copay. Thus we need only consider contracts that offer the lowest copay to the provider(s) charging the lowest price. Now when insurer j offers reduced form contract χ_j, his expected profit per customer is:

$$r(\chi_j) = z_j - (1 - F(\Delta u_j))(\underline{p} - c_j), \qquad \underline{p} = \min(\vec{p}). \tag{11}$$

The insurer's share of individuals (the proportion of individuals who purchase the contract) can also now be defined. When an insurer offers reduced form contract χ_j and the full menu of reduced form contracts is X, that insurer's share of customers is:

$$q(\chi_j|X_{-j}) = \frac{\exp(\overline{u}_j)}{\exp(\overline{u}_j) + \sum_{\ell \neq j} \exp(\overline{u}_\ell)}, \qquad \overline{u}_\ell = \overline{u}(\chi_\ell). \tag{12}$$

This is the well known result of a discrete choice problem with logistically distributed errors. An insurer's profit is simply the product of per customer expected profit and the share of individuals who purchase its contract, conditional on all other reduced form contracts offered:

$$\pi(\chi_j|X_{-j}) = r(\chi_j)q(\chi_j|X_{-j}). \tag{13}$$

While the expected profit from offering a particular contract depends on the contracts of all other insurers, all of the relevant information about these contracts can be summarized by the sum of their exponentiated expected utilities, $\hat{A} = \sum_{\ell \neq j} \exp(\overline{u}(\chi_\ell))$.

Claim 5 *For any set of contracts offered by rival insurers, there exists a (non-negative profit) profit-maximizing contract to offer in response that is almost certainly unique. When it is unique, the best response contract is a continuous function of rivals' contracts.*

It can be shown that there exists a non-negative expected profit best response contract for any menu of rivals' offered contracts X_{-j} and at any price. Moreover, this best response is almost certainly unique,[3] as the insurer's profit function $\pi(\chi_j|X_{-j})$ is single peaked and quasi-concave among contracts that are not useless to individuals. Applying Berge's maximum theorem, by the continuity of the underlying functions and the local quasi-concavity of the profit function, the best response correspondence is a continuous function of $\hat{A}$, the summary statistic of rivals' choices. Proof of the existence of the best response and discussion of its uniqueness are provided in Appendices A.1 and A.3.

3.2.2 Symmetric Equilibrium Among Insurers

Claim 6 *There is a unique and symmetric equilibrium to the subgame among insurers, with all insurers offering the same contract. At any fixed $\vec{p}$, the expected utility of the equilibrium contract is increasing in the number of insurers.*

Note that when the best response to any menu of rivals' contracts (summarized by $\hat{A}$) is a unique contract, the graph of all contracts that are best responses to some menu is a single locus. The monotonicity of the best response functions for the premium and copay are established in Appendix A.4; they are plotted in Figures 3 and 4 respectively. For mathematical convenience, define $A = \sum_{\ell \neq j,0} \exp(\overline{u}(\chi_\ell))$, the same as $\hat{A}$ but now excluding the null contract $(0, \underline{p})$. Labeling these functions as $\check{z}(A)$ and $\check{c}(A)$, we can define a new best response function $\nu_{\underline{p}}(A) = \exp(\overline{u}(\check{z}(A), \check{c}(A))$, the exponentiated expected utility of the best response contract when the sum of exponentiated expected utilities of rival insurers' contracts is A and the price is $\underline{p}$. As both best response functions, the expected contract utility function, and the exponential function are all strictly monotone up, $\nu_{\underline{p}}(\cdot)$ is also strictly increasing.

If there are N insurers, then a symmetric equilibrium occurs at an exponentiated expected utility level $\tilde{u}$ where:

$$\tilde{u} = \nu_{\underline{p}}((N-1)\tilde{u}). \tag{14}$$

That is, if each of insurer j's $N-1$ rivals offers a contract with exponentiated expected utility $\tilde{u}$, then insurer j's best response is to also offer a contract with the same exponentiated expected utility. This best response is unique, thus defining a symmetric equilibrium when all insurers obey the best

[3] Unfortunately, the typical methods for demonstrating that there is only one local maximum of a function are not fruitful in this case. The profit function is not strictly concave everywhere, nor is it even quasi-concave; proof of a single-crossing property of the isoquants of the function's partial derivatives is also elusive. Figure 1 shows a typical contour plot of insurer profit by contract offered. The patterns shown are consistent across parameterizations.

response functions. Appendix B establishes that only symmetric equilibria can exist, that there is at least one symmetric equilibrium, and that the symmetric equilibrium is almost surely unique. Note that the fixed point best response function implicitly depends on the lowest price of care offered by providers, $\underline{p}$. There will be a different symmetric equilibrium among insurers for each price level providers can offer.

The fixed point equation in (14) is demonstrated graphically in Figure 2. The best response function $\tilde{u} = \nu(A)$ is plotted in black versus the sum of exponentiated expected utilities of rival insurers' contracts A, while the fixed point lines for different number of insurers are the colored rays emanating from the origin. For each number of insurers N, the fixed point line is characterized by $\tilde{u} = A/(N-1)$, or an infinitely sloped line when $N = 1$. Equilibrium for each N occurs at the level of A where the best response function crosses the fixed point line, denoted by appropriate colored dots. These levels of A are likewise marked on Figures 3 and 4 to show the equilibrium premium and copay that emerge for each number of insurers. Figure 5 plots the equilibrium contract by number of insurers, independent of the best response functions. Individuals' average expected utility by number of insurers is plotted in Figure 6, again at the baseline parameters in Table 1 with price fixed.[4] Notably, the arrival of the first insurer barely improves expected utility, as a monopolist can offer a contract barely better than no insurance and still capture half of consumers with high per customer profit. The second insurer forces significant competition for customers, resulting in a large cut to the premium and (to a lesser extent) copay; subsequent insurers also lead to more favorable contracts, but with decreasing returns.

3.2.3 Equilibrium Formalized

The previous subsections established that there is a unique symmetric equilibrium of reduced form contracts for any vector of prices $\vec{p}$ offered by providers, which only depends on the lowest price $\underline{p}$. Label this reduced form contract as $\hat{\chi}^* = (z^*(\underline{p}), c^*(\underline{p}))$. As described above, in equilibrium insurers will offer their lowest copay rate only for providers who offer the lowest price. Both individuals and insurers are indifferent between a contract that offers its lowest copay for just one provider and multiple providers, thus there are multiple equilibria in insurance contracts as originally specified when multiple providers tie for the lowest price. For simplicity, temporarily assume that all insurers choose the contract that offers the lowest copay for all providers who bid the lowest price (the "full

[4]Not all individuals have the same expected utility, as preference shocks cause some to choose to be uninsured. Thus the awkwardly phrased "average expected utility".

network"). This will correspond to the "simple equilibrium" in Section 3.3 and will be modified for the "threatening equilibrium". In this way, the equilibrium behavior function for insurers is:

$$\phi(\vec{p}) = (z^*(\underline{p}), c^*(\underline{p}) + \mathbf{1}(\vec{p} > \underline{p})). \tag{15}$$

Defining $\mathbf{1}(\cdot)$ as an indicator function that returns 1 when the statement is true and 0 otherwise (and can operate on and return vectors), equilibrium behavior for all insurers is to offer the equilibrium premium conditional on the lowest price offered by providers. Further, the equilibrium copay is assigned to providers with the lowest price; other providers are assigned a higher copay. This strategy can be seen as representing an insurer's provider network: those who are "in network" are assigned the equilibrium copay, while providers "out of network" are given an arbitrarily higher copay that will never be chosen by individuals.

Claim 7 *There is a unique demand function that maps $\underline{p}$ into the proportion of consumers who will buy medical care when there are N insurers. Total demand $D(\underline{p}, N)$ is strictly decreasing in $\underline{p}$.*

As only the low price $\underline{p}$ matters for the equilibrium contract, demand can be expressed as:

$$D(\underline{p}, N) = \sum_{j=0}^{N} q(\hat{\chi}^* | \hat{X}^*_{-j})(1 - F(\Delta u_j)), \qquad \hat{\chi}_j = \phi(\vec{p}) \text{ for } j > 0, \qquad \hat{\chi}_0 = (0, \vec{p}). \tag{16}$$

Demand for care is the proportion of individuals selecting each contract who purchase care, weighted by the share of each contract. In equilibrium, N insurers will offer the same optimal contract, in addition to the always available null contract. The dependence on N has been implicit throughout this analysis, but is made explicit here to aid in later discussion of how the equilibrium changes with N and M. Demand functions for different number of insurers (at the base parameters) are shown in Figure 9. These functions are strictly decreasing in price, but are neither concave nor convex for their entire spans. Moreover, demand for care is increasing in the number of insurers at higher prices, but is weakly decreasing in N at very low prices. These properties are formally established and discussed in Appendix C.

3.3 Medical Providers' Equilibrium Strategy

If both individuals and insurers obey the equilibrium behavior described above, the game played among medical providers is relatively tractable to analyze. Two equilibria will be presented: first, a

"simple" equilibrium in which the price of care is determined only by providers competing amongst themselves; and second, a "threatening" equilibrium in which insurers use their market power to effectively bargain for a lower equilibrium price than can be achieved in the simple equilibrium. The only difference between the two equilibria is insurers' strategy for including providers in their network (by assigning them the lowest copay). The threatening equilibrium will be primarily used in the analyses in Section 4, with the simple version presented as a benchmark to demonstrate the effect of bargaining. In this section, I drop the N argument from $D(\underline{p}, N)$ for brevity and the lower bar on $\underline{p}$ reduce notational clutter.

Starting with the case where there is only a single provider, we can now fill in (6) as:

$$\hat{\pi}(p) = pD(p) - \kappa(D(p)). \tag{17}$$

Claim 8 *A monopolist medical provider prices above marginal cost, and the equilibrium price does not depend on whether the simple or threatening strategy is used by insurers.*

To maximize profit, the monopolist provider must simply satisfy its first order condition:

$$D(p) = (\kappa'(D(p)) - p)D'(p). \tag{18}$$

The cost function $\kappa(\cdot)$ will be parameterized as a single quadratic term, so that $\kappa'(D(p))$ is linear in demand. The demand function is neither strictly concave nor convex, preventing a mathematical proof that there is a unique solution to the first order condition. In practice, however, the monopolist profit is concave with respect to price and thus has a unique maximizer. The equilibrium when $M = 1$ does not depend on whether the simple or threatening equilibrium is used, as a monopolist cannot be credibly threatened with loss of business. The monopolist prices above marginal cost, as the first factor of the RHS of (18) must be negative for the first order condition to be met (as $D(p)$ is positive while $D'(p)$ is negative).

3.3.1 Simple Equilibrium

In the more interesting case where there are $M \geq 2$ medical providers, it is easy to show that the equilibrium must be symmetric and that all providers will earn non-negative profits; a proof is provided in Appendix D.

Claim 9 *When insurers include in their network any provider offering the lowest price, the range*

of symmetric equilibrium prices is given by $\frac{1}{M}\gamma D(p^*) \leq p^* \leq \left(1+\frac{1}{M}\right)\gamma D(p^*)$.

What prices can be supported as a symmetric equilibrium? Suppose all providers are offering p_0 and earning positive profits, and that the cost function is $\kappa(x) = \gamma x^2$ so that marginal cost is linearly increasing, $\kappa'(x) = 2\gamma x$. No provider could profitably deviate by raising his price, as this would result in zero profit. If a provider lowered his price very slightly, he would capture all of the demand but would need to pay much higher production costs. He can profitably reduce price by a tiny amount when:

$$D(p)p - \kappa(D(p)) > \frac{D(p)}{M}p - \kappa\left(\frac{D(p)}{M}\right) \Longrightarrow D(p)p - \gamma D(p)^2 > \frac{D(p)}{M}p - \gamma\frac{D(p)^2}{M^2} \Longrightarrow \tag{19}$$

$$\Longrightarrow p > \left(1+\frac{1}{M}\right)\gamma D(p).$$

Because $D'(p) < 0$, this inequality has exactly one critical point in p. These prices are not an equilibrium because a provider can profitably deviate. At the opposite end, some prices are so low that even if all providers sold at them, demand (and thus costs) would be so high that all would earn negative profits. In that case, a provider can profitably deviate by raising his price to earn zero instead. Negative profits are earned with symmetric pricing when:

$$\frac{D(p)}{M}p - \kappa\left(\frac{D(p)}{M}\right) < 0 \Longrightarrow p < \frac{1}{M}\gamma D(p). \tag{20}$$

As before, this inequality has a single critical point in p, so the prices ruled out as equilibria are a convex set.

Combining (19) and (20), the prices that constitute symmetric equilibria are given by:

$$\frac{1}{M}\gamma D(p^*) \leq p^* \leq \left(1+\frac{1}{M}\right)\gamma D(p^*). \tag{21}$$

Thus the pure strategy[5] subgame perfect Nash equilibrium with a complete network is not unique, as an entire range of prices are admissible as equilibria. With pure strategies, insurers are unable to wield their market power to encourage providers to lower their prices. If a provider reduces his price from a symmetric equilibrium, insurers will reduce the copay for that provider, pushing all customers to him and thus forcing large production costs relative to having demand split among all

[5]When labeling this a "pure strategy" equilibrium, I ignore the randomization by individuals between providers offering the same price, which cannot be avoided.

providers. In this way, if providers are already pricing at the ceiling of the equilibrium range, they will not compete and arrive at lower equilibrium prices.

3.3.2 Threatening Equilibrium

A key feature of the model is that individuals and insurers are indifferent among providers, conditional on price. If an insurer offers its lowest copay for one provider with the lowest price, the insurer has nothing to gain or lose from also offering the lowest copay for another provider at the low price. Insurers can use this to break the symmetric equilibrium in the upper range of prices given in (21), lowering the ceiling on equilibrium prices. Only a weak form of mixing is required to execute the strategy, and only in a very particular circumstance: when all providers offer the same price that would be an equilibrium under pure strategies, each insurer will randomly choose exactly one provider to cover with the equilibrium copay rather than covering all of them.

Claim 10 *If each insurer randomly selects one provider offering the lowest price to include in its coverage network, demand for each provider's services will have the same mean as in the simple equilibrium, but with positive variance. This reduces expected provider profit while leaving insurer profit unchanged.*

Rather than play the pure strategy given by (15), insurers instead play the mixed strategy:

$$\tilde{\phi}(\vec{p}) = \begin{cases} (z^*(\underline{p}), c^*(\underline{p}) + 1 - e_M(\text{Rand}(\{1, \cdots, M\}))) & \text{if } \overline{p} = \underline{p} > p^{**} \\ (z^*(\underline{p}), c^*(\underline{p}) + \mathbf{1}(\vec{p} > \underline{p})) & \text{otherwise} \end{cases}, \quad \overline{p} = \max(\vec{p}). \tag{22}$$

Here, $e_L(\ell)$ is a vector of length L that has zeros at all indices except ℓ, which has a one. Thus the new strategy calls for insurers to provide the equilibrium copay only to a single, randomly chosen provider (independent across insurers) when providers all offer the same price. This alternative is bounded below by some price p^{**} at which a provider's profit from taking all demand equals the expected profit from the random demand. Subgame perfection is maintained as insurers always offer the equilibrium reduced form contract, but strategically withhold coverage from providers.

When insurers employ the strategy given in (22) and providers price equally, the variance of demand for a particular provider's services takes a positive value, but the mean of demand is unchanged from the pure strategy equilibrium– each provider will receive $\frac{1}{M}$ of the total demand $D(p)$ on average. With the price of care fixed at some p, each provider faces uncertain profit. When

marginal cost is linear (as above), profit follows a quadratic form with an expected value of:

$$\mathrm{E}[\hat{\pi}_k|\underline{p} = \overline{p}] = \frac{D(p)}{M}p - \gamma\left(\varsigma^2 + \left(\frac{D(p)}{M}\right)^2\right), \quad \varsigma^2 \approx \frac{M-1}{M^2N}D(p)^2. \tag{23}$$

In this equation, ς^2 is defined as the variance of demand for provider k's services. When the mixed strategy is (22), demand is the sum of N Bernoulli random variables, each returning $\frac{1}{N}$ with probability $\frac{1}{M}$ and zero otherwise, thus the form of ς^2 above.[6]

Claim 11 *When each insurer randomly chooses one provider to include in their coverage network if all providers price equally, the ceiling on equilibrium prices is given by* $p^{**} = \left(1 + \frac{1}{M} - \frac{1}{MN}\right)\gamma D(p^{**})$. *This ceiling is lower than when insurers use the simple strategy.*

If a provider can choose a p very slightly below the current symmetric pricing and achieve higher expected profit, then the current $\underline{p}$ is not an equilibrium under mixed insurer strategies. Combining the two parts of (23), this occurs when:

$$D(p)p - \gamma D(p)^2 > \frac{D(p)}{M}p - \left(\frac{M-1}{M^2N} + \frac{1}{M^2}\right)\gamma D(p)^2 \Longrightarrow p > \left(1 + \frac{1}{M} - \frac{1}{MN}\right)D(p). \tag{24}$$

The unique critical point of this inequality defines p^{**}, the ceiling of the equilibrium price range when insurers use mixed strategies. If providers are pricing symmetrically at or below p^{**}, the threat of facing random demand does not induce them to undercut their competitors, as absorbing all of the demand would generate such large costs. Thus the new range of equilibrium prices is characterized by:

$$\frac{1}{M}\gamma D(p^*) \leq p^* \leq \left(1 + \frac{1}{M} - \frac{1}{MN}\right)\gamma D(p^*). \tag{25}$$

Allowing mixed strategies among insurers lowers the ceiling of equilibrium prices, as the rightmost term of (25) is smaller than that of (21). The floor is unchanged, as providers will never accept symmetric pricing that results in negative expected profits, preferring to price arbitrarily high and guarantee zero profit. It is relatively clear that, as in the pure strategy equilibrium, the ceiling is decreasing in the number of medical providers due to competitive effects. More importantly, the upper bound of equilibrium prices now depends on the number of insurers. When an additional

[6]The definition of ς^2 is actually an approximation. So long as a provider is priced at the lowest price among rivals, demand for its services can never fall to zero. Even if no insurer randomly selects provider k for the lowest copay, k will still have uninsured customers. The approximation is very accurate for several reasons: the share of uninsured customers tends to be small, the proportion of uninsured individuals who buy care is small, and the probability that no insurer covers a particular provider is small when $N > 2$.

insurer enters the market, the variance of demand decreases, reducing the threat value of the random demand faced by providers. Thus the ceiling of equilibrium prices is increasing in the number of insurers due to reduced bargaining or threat power.

Claim 12 *The symmetric equilibrium price can be expressed as the market clearing price between demand for care $D(\underline{p}; N)$, determined by the equilibrium of the subgame among insurers, and a "pseudo-supply" function $S(p; M, N)$, determined by the equilibrium among providers given in* (25)*.*

The right half of (25) can be rearranged to express the equilibrium price ceiling as a "pseudo supply curve", yielding:

$$S(p; M, N) = \frac{p}{\gamma\left(1 + \frac{1}{M} - \frac{1}{MN}\right)} \text{ for } N > 0 \;\&\; M > 1, \qquad S(p; M, N) = \frac{p}{\gamma\left(1 + \frac{1}{M}\right)} \text{ for } N = 0. \quad (26)$$

When there are no insurers or only one medical provider, the threatening equilibrium cannot be used; the right half of (21) is used when there are no insurers, while this analysis does not apply at all when there is a monopolist provider. This linear function of the price of care represents the combinations of price and quantity of care that could be supply-side (ceiling) equilibria, rather than the quantity of care that firms are willing to supply at each price as in a traditional supply curve. The price that satisfies $S(p; M, N) = D(p; N)$ is thus the unique equilibrium price that will be offered by all medical providers. The determination of equilibrium price and quantity is shown in Figure 10, when there are two medical providers and zero to nine insurers; the supply and demand curves share a color for each number of insurers, and the relevant intersections are marked with appropriately colored dots.

Claim 13 *The ceiling of the equilibrium price range is never less than marginal cost at the quantity of care produced. Price equals marginal cost iff there is exactly one insurer and two providers.*

As a benchmark comparison, the top end of the equilibrium price range is never less than the competitive level where price equals marginal cost. If all medical providers price equally, the competitive price is characterized by $\frac{2}{M}\gamma D(p) = p$. When there is only a single insurer and exactly two medical providers, the monopolist insurer can force the price to the competitive level, but at any other combination of insurers and providers the top end of the equilibrium range will be higher. Just as in the subgame among insurers, the model is characterized by imperfect competition among providers in nearly all cases. As a final note, the mixed strategy employed by insurers is ideal

from the perspective of individuals, as it maximizes the threat power of insurers (while maintaining subgame perfection) and thus results in the lowest equilibrium price ceiling possible.

4 Competition and Consumer Welfare

Having characterized equilibrium behavior for all players, we can now consider how competition among insurers and medical providers affects outcomes. The utility function is specified as CRRA, while the distribution of medical needs is exponential with mean λ; baseline parameters are provided in Table 1. The analysis begins with equilibrium outcomes from various numbers of insurers and medical providers for an example parameter set. The two objects of interest are the top end of the equilibrium price of care and the certainty equivalent level of consumption. Additional simulations are also presented to demonstrate that some results of the main analysis are not universal, but depend on the parameters chosen. In Section 4.3, I decompose the effects of insurer entry, separating changes due to increased competition, shifts in demand, and loss of bargaining power by selectively turning off these channels. Finally, an alternative specification in which medical needs are observable (eliminating moral hazard) is presented and briefly discussed.

Somewhat unusually, the model yields a range of equilibrium prices of medical care rather than a unique value. The examples below assume that the ceiling of the equilibrium range is the price that will emerge, a choice validated by the structure of the game. From any configuration of prices offered by medical providers that are all greater than the equilibrium ceiling, a sequence of best responses to rivals' current choices will converge to the equilibrium ceiling.[7] In this way, the top of the equilibrium price range is a "stable sink" for a competitive game among providers, while lower prices cannot be reached from above. Further justification for this (trivial) equilibrium selection mechanism is that the ceiling of the equilibrium price range yields the highest profit for medical providers– it is the optimal equilibrium for the players who get to move first and "lead the dance". Most critically, the equilibrium ceiling varies with the number of both insurers and providers, allowing analysis with respect to competition.

The measure of consumer welfare used in the examples computes the certainty equivalent level of consumption compared to a baseline with no medical care available. Even with insurance, individuals

[7] Technical caveat: The space of prices must be finitely discretized (e.g. to the penny) for the sequence to converge, otherwise successive best responses would move by infinitesimal amounts. The convergence can also occur for configurations where some (but not all) providers begin priced below the equilibrium ceiling, depending on the order of moves.

face utility risk through both consumption (due to a copay) and untreated medical needs. If the expected utility across individuals, appropriately weighting by insured and uninsured, is run through the inverse of the utility function, it yields the certainty equivalent level of consumption. The ratio of this value to the certainty equivalent when individuals cannot purchase medical care (less one) is a normalized metric of the relative value of a competitive scenario: the percentage increase in certainty equivalent consumption that the insurers and medical providers give to individuals. Preference shocks are not included in the calculation of expected utility when computing the welfare metric, as these shocks are meant to represent imperfections in individuals' understanding of or access to contracts and thus should not be considered boons to welfare. Moreover, the average preference shock of the selected insurance contract does not vary much once there are at least two insurers, so there is very little differential effect on welfare.[8]

4.1 Number of Insurers

When an additional insurer enters the market, the welfare of individuals is affected in three ways. First, at any given price of care that providers offer, there will be more competition among insurers, resulting in a more favorable equilibrium contract and increasing expected utility (the "competition effect"). Second, the change in equilibrium contract (as well as a higher proportion of insured individuals merely due to the arrival of a new option) at every price shifts the total demand for care function from $D(p, N)$ to $D(p, N+1)$, altering the equilibrium price range defined by (25) (the "demand effect"); the direction of this effect is ambiguous, but reduces the expected utility of individuals in the typical case when total demand increases. Third, the additional insurer dampens the threat power of insurers against providers by reducing the variance of demand they can induce, increasing the equilibrium price and thus reducing expected utility (the "bargaining effect"). The competition effect benefits individuals directly, while the demand (usually) and bargaining effects harm them indirectly through a higher price of care. Note that each individual would like to be offered the more favorable contract from increased competition, but would prefer if everyone else was not– the collective increase in demand harms all consumers, even though it arises from individually beneficial choices.

The examples below demonstrates that the latter two effects can sometimes dominate the competition effect, so that increasing the number of insurers reduces average individual welfare. In

[8]The welfare values could be expressed as raw utilities, but these would be difficult to read. With $\rho = 5$, all expected utility values are negative and very close to zero.

contrast, increasing the number of medical providers always strictly improves individuals' expected utility, as the only effect is to reduce the equilibrium price of care. For each combination of N and M, the model is solved with the parameter values listed in Table 1. The equilibrium price of care and the certainty equivalent level of consumption at each combination are presented in Tables 2 and 3 respectively.

Example 1 *The entry of each insurer beyond the first always increases the equilibrium price of care as the demand curve shifts outward and threat power is lost. The arrival of the first insurer can sometimes reduce equilibrium price if the bargaining effect overcomes the demand effect.*

As expected, the equilibrium price of medical care increases with additional insurers (reading down any column of Table 2) due to both the greater total demand for care and the reduced threat power of insurers. The only exception to this pattern is when the first insurer enters the market when there are two, three, or four medical providers. When there is only one provider, the first insurer has no ability to threaten the monopolist provider, thus the equilibrium price increases due to greater demand for care. When there are many providers, the first insurer has threat power to reduce the price of care, but this effect is overwhelmed by the increase in demand. Only at "moderate" levels of providers can threat power dominate the countervailing effect. As noted above, the equilibrium price decreases with the number of medical providers (reading across any row); the only exception is when there is exactly one insurer and at least two providers, when the price is constant (see (25) when $N = 1$). Note in particular the relatively large drop in the equilibrium price when the second medical provider enters the market. Rather than freely choosing the price of care to maximize profit, the providers must compete with each other. Moreover, they are susceptible to the threatening strategy of insurers, unlike the monopolist, further reducing the equilibrium price.

Example 2 *Consumer welfare can follow a U-shaped pattern with respect to additional insurers, falling as the first few insurers enter and rising with subsequent entry.*

Reading down any column of Table 3 reveals that the effect of entry on consumer welfare is not monotone. Instead, welfare seems to follow a U-shape as insurers enter the market. For example, when there are two medical providers, welfare is 7.56% above baseline with one insurer, falls to 4.28% after the second and third insurers enter the market, but then steadily increases to 4.37% when nine insurers have entered. All other columns follow the same pattern, with the "welfare trough" consistently occurring when there are three insurers. As seen in Figure 7, the boost in

the total demand function is quite large when an insurer enters and there are few incumbents, but subsequent entry leads to successively smaller changes.[9] Moreover, the loss in threat power from an additional insurer becomes smaller as the number of incumbent insurers increases. Thus the demand and bargaining effects of entry are strongest when N is small, tapering off to be dominated by the competition effect as N grows, generating the U-shaped pattern seen in each column.

Example 3 *It is possible for consumers to be better off with no insurance available than with any number of insurers.*

Most strikingly, the adverse effect of an additional insurer is strongest when moving from zero to one insurers: welfare drops when moving from the first to the second row of Table 3. The shift in the demand function (and thus in price) is so great that it more than offsets the consumption-smoothing benefits of insurance and the ability for the insurer to use its threat power to negotiate a lower equilibrium price. Moreover, a monopolist insurer need not offer a favorable contract due to lack of competition. The only counterexample occurs when there are exactly two medical providers. More strongly, this is the only case where the presence of *any* number of insurers is able to increase consumer welfare above the level it reaches in the absence of insurers. This surprising and counter-intuitive result is not universal, however; see Section 4.3 for an alternative parameterization where the presence of insurers does raise individuals' expected utility. Finally, note that increasing the number of medical providers strictly increases welfare through increased price competition among providers (except when there is exactly one insurer).

In the baseline example, individuals' welfare is almost never higher in the presence of insurers– that the maximum welfare value tends to occur in the top row, with zero insurers. This pattern is particular to the parameters chosen and does not hold universally.

Example 4 *It is possible for the entry of each insurer beyond the first to strictly increase consumer welfare if medical needs are sufficiently large.*

Table 4 shows the relative welfare gains for each combination of insurers and medical providers when medical needs shocks are twice as large as in the benchmark example ($\lambda = 0.5$). In this case, the entry of an additional insurer (as long as there already at least one) is always strictly welfare-improving, with no U-shape in any column. Moreover, as more insurers are added in each column, individuals' welfare eventually overtakes its original level with no insurers. The critical number of

[9] Figure 7 holds the price of care at a fixed level, which is not the case in the full welfare analysis. Similar logic applies, even in the presence of other effects.

insurers at which this occurs is greater when there are more medical providers, as the welfare loss from the entry of the first insurer becomes greater and greater. Note that the values in Table 4 are not comparable to those in any other table, as the change in λ lowers the denominator of the certainty equivalent consumption through higher average medical needs.

4.2 Variance of Preference Shocks

As mentioned in Section 2.6, the preference shocks over insurance contracts can be interpreted as informational barriers to perfect competition among insurers. They could represent difficulty in understanding and comparing contracts (where the simple model is a stand-in for complex real world contracts), search costs, or the effect of purchasing insurance through an employer. This section considers how the variance of the preference shocks affects individuals' welfare. Intuitively, increasing the magnitude of the preference shocks relative to the utility function should usually harm individuals, inducing them to "make mistakes" more often and allowing insurers to offer less favorable contracts, knowing that many individuals will choose them due to a large preference shock. While this intuition usually holds, it is not a universal result.

Example 5 *When individuals have no preference shocks over insurance contracts so there is perfect competition among two or more insurers, the equilibrium price and quantity of care are much higher because individuals are offered very favorable contracts; consumer welfare is much higher than in the presence of preference shocks. When there is a monopolist insurer and no preference shocks, consumer welfare is much lower than under imperfect competition.*

In the extreme case where preference shocks are eliminated entirely, competition from two or more insurers pushes the equilibrium contract to the expected utility-maximizing point on the zero profit locus (the "perfect competition contract"). In this case, all consumers will choose the best contract among their choices, without preference error. Suppose an insurer offers the current best contract and it is not the perfect competition contract.[10] Then another firm could choose a nearby contract that offered slightly more favorable terms to individuals while still yielding positive profit, capturing the entire market. The zero profit portion of the perfect competition result can also be seen as the limit of (44) as the scale of Δu approaches infinity and both terms go to zero. The equilibrium contract will vary with the price of care, as the insurer must break even to remain

[10] As before, this contract must yield non-negative profit, else the insurer could improve its payoff by choosing a different contract. See Appendix F for derivations.

in business, generating a "perfect competition demand curve", plotted on Figure 12.[11] Under the additional assumption that these insurers collude against providers during negotiations (or there is a benevolent monopolist insurer), the perfect competition equilibrium can be determined. With insurer profit eliminated, demand for care under perfect competition is significantly higher as consumers are offered more generous contracts. In equilibrium, both the price of care and total quantity of care are much higher than in the imperfectly competitive scenarios presented in Tables 2 and 3. However, welfare is also significantly higher, achieving a level about twice as high as the gain from most combinations of insurers and providers; indeed, consumers' welfare is even higher than in the most favorable scenario considered in Section 4.3, when the demand effect is turned off.

In contrast, the absence of preference shocks can result less favorable contracts when there is only one insurer. In this case, the monopolist insurer can capture the entire market as long as he offers a contract that is better than the null contract. He will thus choose the profit-maximizing contract on the same indifference curve as the null contract (or a contract arbitrarily close to it). As it is possible that a monopolist insurer will offer a contract with greater expected utility than the null contract when individuals do have preference shocks, this means that a larger magnitude of preference shocks is not strictly worse for individuals in all cases. To better understand the ambiguous effects of preference shocks on individuals' welfare in non-extreme cases, specific examples are necessary. Tables 5 and 6 demonstrate the welfare effects when the variance of preference shocks is doubled and halved respectively. In each example, the base parameters in Table 1 are used, with only the relative magnitude of preference shocks altered.

Example 6 *It is possible for consumer welfare to be strictly decreasing in the number of insurers when preference shocks over insurance contracts are large.*

In Table 5, where preference shocks are larger, individuals are strictly worse off (relative to their welfare in Table 3), regardless of the number of insurers or medical providers. In contrast to the benchmark scenario, an additional insurer always harms individuals (with one exception), as competition between insurers is much weaker and thus the gains from competition cannot overcome the loss of threat power and the shift in the total demand function. For example, when there are three medical providers, welfare is 7.2% above baseline with one insurer but falls to 4.55% with two insurers, steadily down to 2.71% with nine insurers. Surprisingly, individuals can even be made

[11] With no preference shocks, individuals make no errors when selecting a contract. Because an actuarially fair contract offering positive insurance is definitely better than being uninsured, all individuals buy insurance under perfect competition.

worse off than they are when medical care is unavailable, as in the lower left portion of Table 5. As more insurers enter the market and a larger portion of the population is insured, the high price of care due to the provider's monopoly and shifted demand function more than offsets the improvement from being able to purchase care.

Example 7 *Consumers are better off when preference shocks are smaller as long as there is more than one insurer.*

Table 6 represents the opposite case, where the magnitude of preference shocks is smaller than the benchmark case. The U-shaped pattern of welfare in each column returns, with the trough shifted to the row with two insurers rather than the original three. As expected, individuals' welfare is almost always higher when they face smaller informational barriers to choosing an ideal contract; they are "better shoppers", forcing insurers to compete for their business. The only exception is when there is exactly one insurer and at least two medical providers. As described above, this situation allows the monopolist insurer to only surpass the null contract by a small amount of utility and still capture much of the market to maximize his profit. In this way, reducing imperfections in the market can actually harm individuals in limited circumstances. While the price tables are omitted for space, it is worth noting that when the variance of preference shocks is small the equilibrium price of care is consistently higher than the benchmark experiment. This arises precisely because the equilibrium contract is so favorable to individuals that they are able to purchase care for their medical needs more often. The reverse is true when the preference shock variance is low, so that welfare does not track with the price of care as closely as Tables 2 and 3 seemingly indicate.

Example 8 *Perfect competition does not lead to the consumer-welfare maximizing outcome in equilibrium. A benevolent social planner can offer an insurance contract with higher welfare, at a lower price and demand for care.*

While eliminating preference shocks over insurance contracts vastly improves consumer welfare, the equilibrium of this market is not actually the ideal outcome for consumers. Under perfect competition, individuals collectively "get greedy" and want to purchase large quantities of care. Neither consumers nor insurers take into account the effect their additional demand for care has on the price insurers are willing to sell for. It is thus possible for a social planner to offer an insurance contract with even higher expected utility than achieved under perfect competition by maximizing individuals' utility with the constraint that the contract must result in an outcome where $p = \gamma D(p)$,

on providers' pseudo-supply curve (under a monopolist or collusively bargaining insurers). The social planner thus "slides down" the supply curve to achieve a lower equilibrium price. As shown in Table 12, consumer welfare improves significantly even as the quantity of care purchased falls by nearly 8%. The welfare gains are achieved through significantly higher utility when individuals don't purchase care, due to a 40% lower premium, blunted by somewhat lower utility when they do purchase care (as $z+c$ is larger than under perfect competition). The social planner's demand curve is also plotted on Figure 12, and derivations are provided in Appendix F.

4.3 Decomposition of the Effects of Insurer Entry

As previously discussed, there are three effects from the entry of an additional insurer: consumers are offered more favorable contracts due to increased competition (at any given price), total demand for care at any price increases as a result of these more favorable contracts, and bargaining or threat power is eroded. This section decomposes the total effect of insurer entry on consumer welfare by presenting equilibrium results of alternative games where one or more of these effects are "turned off". In summary, the U-shaped pattern of welfare with respect to the number of insurers can occur even if only the bargaining or demand effect is present to counteract the competitive effect; when both effects are turned off, insurer entry strictly improves consumer welfare.

Example 9 *Consumer welfare can decrease in the number of insurers even if there is no loss of bargaining power from the entry of an additional insurer. This can occur whether bargaining never happens at all (as in the simple equilibrium) or if bargaining power is maintained through insurer collusion against providers.*

The increase in the equilibrium price of care from the shift in demand can more than offset the gains from increased competition among insurers (in the language of Section 4.1, the demand effect can trump the competition effect even without help from the bargaining effect). Table 7 shows individuals' welfare when insurers use the simple equilibrium described in Section 3.3.1 rather than the threatening equilibrium assumed in the other examples, so that the equilibrium price is defined by the top end of (21) instead of (25). Graphically, the equilibria for this scenario can be seen in Figure 10 as the intersection of the light green pseudo-supply curve (for zero insurers, when bargaining cannot occur) with each of the demand curves.[12] That is, when an additional insurer

[12]Figure 10 shows pseudo supply curves for $M = 2$ only; the slope of the light green pseudo-supply curve and all others would be steeper for $M > 2$, but the mechanics remain the same.

enters, we move to a higher demand curve but the pseudo-supply curve remains as if $N = 0$. As in the benchmark example, welfare initially declines as insurers enter the market before increasing at larger numbers; the trough consistently occurs when there are exactly two insurers. While the top row and the leftmost column are identical (as the threatening strategy cannot be used on a monopolist provider), all other values in Table 7 are smaller than their counterparts in Table 3, with the largest losses occurring when there is only one insurer, when threat power would be highest.

In a similar exercise, the rightmost column of Table 7 (labeled "C", for "collusion", "cartel", or "coordination") presents individuals' welfare when insurers maintain the full threat power of a single insurer, regardless of their actual number.[13] This represents a situation in which insurers coordinate their randomization, selecting the same single provider to cover. Turning again to Figure 10, this corresponds to the intersection between the red pseudo-supply curve and each successive demand curve (other than the blue "zero insurers" curve), as all insurers coordinate as single insurer when bargaining for prices. The familiar non-monotone pattern to individuals' welfare is also present here, reinforcing that the demand effect can overcome the competition effect even without the bargaining effect. Comparing the values in column "C" to any column of Table 3 reveals the extent of welfare loss from the bargaining effect, which ranges from 8% to 38% of the total welfare benefit of access to medical care. Predictably, the bargaining effect harms individuals most when there is a large number of insurers and a small number of providers.

Example 10 *Consumer welfare strictly increases with additional insurers beyond the first when the demand curve does not shift outward and there is no loss of bargaining power. Welfare can be non-monotone in the number of insurers when there is no demand effect but bargaining power erodes with insurer entry.*

Table 8 eliminates the demand effect along with the bargaining effect by solving the model as if the total demand curve did not change as insurers enter the market, isolating only the competitive effect. This could represent an odd world in which only a single individual is offered insurance by potentially multiple firms, allowing him to experience only the benefits of the competition effect, without either countervailing effect.[14] Graphically, we change neither the demand curve nor the pseudo-supply curve, remaining permanently pinned at the intersection of the blue lines; price is

[13] Only a single column is needed, as the equilibrium price does not change with the number of providers, as long as there are at least two so that the threatening strategy can be used. The top row is omitted because there are no insurers and thus no threat power, while the row with a single insurer is obviously identical to its benchmark counterpart.

[14] The threatening strategy cannot be used, as insurers each have a measure zero set of customers. The welfare table is shown from the perspective of the lucky individual being offered insurance.

identical for any number of insurers. Comparing values in Table 8 to their corresponding entries in Table 7 reveals that upwards of 50% of the potential welfare gains from access to medical care are lost to the demand effect. From a political perspective, this could be informative as to why already insured individuals may oppose reforms to expand access to insurance, fearing a price increase from the demand effect. Note that individual welfare is strictly increasing in the number of insurers (after the first one), quickly overtaking the values in the top row in the absence of insurance. Welfare drops slightly when the first insurer enters, as in this case a monopolist insurer will offer a slightly worse contract than the null contract. Finally, Table 9 presents welfare results from an impossible world in which there is no demand effect from additional insurers, but insurers maintain bargaining or threat power in determining the equilibrium price of care. The U-shaped pattern returns, with welfare falling as the second insurer enters (when there are fewer than seven providers).

4.4 Observable Medical Needs

The demand effect arises from moral hazard: as individuals are offered more favorable contracts, they purchase medical care a greater portion of the time (they have a lower critical level of medical need) thus increasing total demand for care and the equilibrium price. Similarly, moral hazard harms individuals as it prevents insurers from offering complete insurance– a copayment is necessary to prevent individuals from purchasing care 100% of the time. As an alternative model, Table 10 presents welfare results from a world in which this problem does not occur because medical needs are observable and contractible by insurers. In this world (whose details can be found in Appendix E), insurance contracts specify a premium and a threshold medical need level, above which individuals are entitled to medical care at no out-of-pocket cost.

Example 11 *When medical needs are observable and contractible, equilibrium insurance contracts are complete: individuals are not charged a copay, but instead specify a threshold level of medical need for care to be fully reimbursed. At any positive number of insurers and medical providers, consumer welfare is higher than when medical needs are private information. When there are at least two providers, equilibrium prices are lower than the baseline model because moral hazard is contained.*

These welfare values are clearly much higher than any other scenario presented, as insurers can offer much more favorable contracts when they can control the rate at which their customers will purchase care; moreover, individuals face no consumption risk, further boosting welfare. Notably, the competition effect consistently dominates the demand and bargaining effects as additional insurers

enter the market, in contrast to the baseline scenario. As shown in Table 11, part of the welfare gains are due to the lower equilibrium price of care as compared to the baseline in Table 2; in nearly all columns, equilibrium price is lower than in the baseline model when there is only one insurer, and grows more slowly than in Table 2 as insurers enter. The only exception is when there is a monopolist medical provider, and even then welfare is still much higher under observable medical needs due to complete insurance.[15] While perfect observability of the health shock is an extreme assumption, the example illustrates how insurers' efforts to monitor customers' need for particular services might actually benefit consumers ex ante.

Just as in the baseline model, we can compute the perfect competition demand curve when medical needs are observable rather than private information. In this case, both demand and price are even lower than the social planner's solution when medical needs are not contractible, while premiums are much higher (as expected, as it is the only source of insurers' revenue). Even with slightly lower demand (and thus individuals' facing the pain cost of their medical conditions more often), consumer welfare is more than double that of either "ideal outcome" under private information, and much higher than the imperfectly competitive equilibria presented in Table 10. With observable medical need and complete contracts, there is no loss of consumption utility when care is purchased. Thus the insurer is able to fully complete both of its welfare enhancing roles: allowing individuals to purchase care more often (reducing the probability and magnitude of losses) and eliminating consumption risk. Once again, the perfect competition outcome can be improved by a social planner who can commit to offering a contract that isn't individuals' ideal outcome at a particular price, but is the ideal contract given providers' constraint. As seen in the baseline private information model, the social planner's solution slides down the pseudo-supply curve, reducing demand by 12% as premiums fall by over 22% due to the lower price achieved (note that $z = pD$ with a zero profit condition and observable medical needs). Consumer welfare increases slightly, but proportionately less than in the baseline model.

[15] Prices under a monopolist medical provider are much higher under observable medical needs than in the baseline scenario because of the slope of the demand curves; see Figure 11. When there is at least one insurer, the demand curve is much steeper than baseline at low prices, as individuals get benefits from lower consumption risk even as they buy care less often. Demand levels off and is very inelastic at higher prices. The profit maximizing monopolist provider is thus free to raise price without sacrificing much quantity. Note than equilibrium price exceeds individuals' income ($y = 2$) in most cases, so that uninsured individuals cannot buy care at all.

5 Conclusion

This chapter presents a three-way model of the medical care market among individuals, insurers, and providers. The model is solved for a range of subgame perfect equilibria prices, where one endpoint of the range is a stable sink for best responses to non-equilibrium strategies and thus the natural choice for the equilibrium that will actually occur. Both analytically and through examples, the relationship between the number of insurers and individuals' welfare is shown to be not strictly monotonic, instead following a U-shaped pattern as insurers enter the market.

The pattern occurs because there are three effects that arise from the entry of an additional insurer. First, an additional insurer creates additional competition among insurers, resulting in more favorable contracts being offered to individuals conditional on the price of care set by providers. Second, the new contract changes the probability that an insured individual will purchase care, (usually) increasing the total demand for care conditional on price and thus increasing the equilibrium price. Third, the new insurer cannibalizes the bargaining or threat power of existing insurers when determining the equilibrium price of care. The demand and bargaining effects can overwhelm the competition effect, leading to a net welfare loss; more strongly, either the demand or bargaining effect can be sufficient on its own to overcome the competition effect.

The model is the most simple specification that displays the complex relationship between insurer competition and consumer welfare. Notably, all individuals are ex ante identical, with no differences in medical risk, preferences, or resources; insurers and medical providers have similar parity, with no shifters for cost or quality. While previous literature has included these features, they are omitted here to focus on competition among insurers and providers and its effect on welfare. Further, the "bargaining" that occurs between insurers and providers is admittedly rudimentary, but it allows for the dilution of market power with additional entry, the critical result of more complex mechanisms such as Nash bargaining. Finally, the model parameters used in the examples are not calibrated to match real world outcomes, merely selected to demonstrate the counterintuitive results of the model. On the whole, the model should be considered as demonstrative rather than predictive.

With the mixed empirical evidence on how the concentration of insurer market power affects consumer welfare and the recent introduction of insurance exchanges through the Affordable Care Act, a more complete model with some of the above elements is perhaps warranted. Given appropriate data, the estimated model could be informative about whether competitive reforms will harm rather than help consumers. In particular, a version with continuous (rather than unit) medical

care can address the issue of partial observability of medical needs, a major driver of lost potential welfare. Alternatively, including individuals with heterogeneous characteristics can reveal how competitive reforms might not have uniform effects across the population. These avenues serve as future directions for continuing this line of research.

References

Bamezai, A., Zwanziger, J., Melnick, G. A., and Mann, J. M. (1999). "Price Competition and Hospital Cost Growth in the United States (1989-1994)." *Health Economics*, 8: 233–243.

Bates, L. J. and Santerre, R. E. (2008). "Do Health Insurers Possess Monopsony Power in the Hospital Services Industry?" *International Journal of Health Care Finance Economics*, 8: 1–11.

Caplin, A. and Nalebuff, B. (1991). "Aggregation and Imperfect Competition: On the Existence of Equilibrium." *Econometrica*, 59(1): 25–29.

Capps, C. and Dranove, D. (2004). "Hospital Negotiation and Negotiated PPO Prices." *Health Affairs*, 23(2): 175–181.

Capps, C., Dranove, D., and Satterthwaite, M. (2003). "Competition and Market Power in Option Demand Markets." *RAND Journal of Economics*, 34(4): 737–763.

Dafny, L., Duggan, M., and Ramanarayanan, S. (2009). "Paying a Premium on Your Premium? Consolidation in the U.S. Health Insurance Industry." *NBER Working Paper*, (15434).

Dafny, L. S. (2010). "Are Health Insurance Markets Competitive?" *American Economic Review*, 100: 1399–1431.

Frank, R. G. and Lamiraud, K. (2009). "Choice, price competition and complexity in markets for health insurance." *Journal of Economic Behavior & Organization*, 71: 550–562.

Gal-Or, E. (1997). "Exclusionary Equilibria in Health-Care Markets." *Journal of Economics & Management Strategy*, 6(1): 5–43.

Gaynor, M. and Town, R. J. (2011). "Competition in Health Care markets." *NBER Working Paper*, (17208).

Gaynor, M. and Vogt, W. B. (2003). "Competition Among Hospitals." *RAND Journal of Economics*, 34(4): 764–785.

Gaynor, M., Ho, K., and Town, R. (2014). "The Industrial Organization of Health Care Markets." *NBER Working Paper*, (19800).

Haas-Wilson, D. and Garmon, C. (2011). "Hospital Mergers and Competitive Effects: Two Retrospective Analyses." *International Journal of the Economics of Business*, 18(1): 17–32.

Ho, K. (2009). "Insurer-Provider Networks in the Medical Care Market." *American Economic Review*, 99(1): 393–430.

Ho, K. and Lee, R. S. (2013). "Insurer Competition and Negotiated Hospital Prices." *NBER Working Paper*, (19401).

Inderst, R. and Wey, C. (2003). "Bargaining, mergers, and technology choice in bilaterally oligopolistic industries." *RAND Journal of Economics*, 34(1): 1–19.

Maestas, N., Shroeder, M., and Goldman, D. (2009). "Price Variation in Markets with Homogeneous Goods: The Case of Medigap." *NBER Working Paper*, (14679).

Melnick, G. A., Zwanziger, J., Bamezai, A., and Pattison, R. (1992). "The Effects of Market Structure and Bargaining Position on Hospital Prices." *Journal of Health Economics*, 11: 217–233.

Melnick, G. A., Shen, Y.-C., and Wu, V. Y. (2011). "The Increased Concentration of Health Plan Markets Can Benefit Consumers Through Lower Hospital Prices." *Health Affairs*, 30(9): 1728–1733.

Rochet, J.-C. and Tirole, J. (2004). "Two-Sided Markets: An Overview." *FRB Atlanta working paper*.

Staten, M., Dunkelberg, W., and Umbeck, J. (1987). "Market Share and the Illusion of Power." *Journal of Health Economics*, 6: 43–58.

Trish, E. E. and Herring, B. J. (2014). "How Do Health Insurer Market Concentration and Bargaining Power with Hospitals Affect Health Insurance Premiums?" *Working paper*.

Wu, V. Y. (2009). "Managed Care's Price Bargaining With Hospitals." *Journal of Health Economics*, 28: 35–360.

Table 1: Baseline Values of Parameters Used In Examples

Parameter	Value	Description
y	2	Individuals' income
ρ	5	Coefficient of relative risk aversion
λ	0.25	Mean of medical needs distribution
σ	0.05	Scaling factor for preference shocks
γ	3	Slope of marginal cost of producing care
p	1.2	Fixed price of care (where relevant)

Table 2: Equilibrium Price of Care by Number of Insurers and Medical Providers, Baseline

	# of Medical Providers								
# of Insurers	1	2	3	4	5	6	7	8	9
0	1.140	1.092	1.076	1.066	1.060	1.056	1.052	1.050	1.048
1	1.276	1.065	1.065	1.065	1.065	1.065	1.065	1.065	1.065
2	1.534	1.195	1.170	1.158	1.150	1.144	1.141	1.138	1.135
3	1.632	1.285	1.248	1.228	1.215	1.207	1.200	1.196	1.192
4	1.687	1.335	1.292	1.269	1.254	1.244	1.237	1.231	1.227
5	1.723	1.365	1.319	1.294	1.278	1.267	1.259	1.253	1.248
6	1.748	1.385	1.337	1.310	1.293	1.282	1.273	1.267	1.262
7	1.768	1.400	1.349	1.322	1.304	1.292	1.284	1.277	1.272
8	1.783	1.410	1.359	1.330	1.313	1.300	1.291	1.285	1.279
9	1.795	1.419	1.366	1.337	1.319	1.307	1.297	1.290	1.285

Table 3: Individuals' Welfare by Number of Insurers and Medical Providers, Baseline

	# of Medical Providers								
# of Insurers	1	2	3	4	5	6	7	8	9
0	4.51%	6.75%	7.62%	8.14%	8.49%	8.74%	8.93%	9.08%	9.20%
1	1.08%	7.56%	7.56%	7.56%	7.56%	7.56%	7.56%	7.56%	7.56%
2	0.84%	4.56%	5.13%	5.45%	5.66%	5.80%	5.91%	5.99%	6.05%
3	0.97%	4.28%	4.86%	5.21%	5.43%	5.59%	5.71%	5.81%	5.88%
4	1.02%	4.30%	4.91%	5.26%	5.50%	5.66%	5.79%	5.89%	5.96%
5	1.04%	4.32%	4.95%	5.33%	5.57%	5.74%	5.87%	5.97%	6.05%
6	1.05%	4.34%	4.99%	5.37%	5.62%	5.80%	5.94%	6.04%	6.12%
7	1.05%	4.35%	5.02%	5.41%	5.67%	5.85%	5.98%	6.09%	6.17%
8	1.05%	4.36%	5.04%	5.43%	5.70%	5.88%	6.02%	6.13%	6.21%
9	1.05%	4.37%	5.05%	5.45%	5.72%	5.91%	6.05%	6.16%	6.25%

Note: Tables 3-10 present individuals' welfare as the percentage change in certainty equivalent consumption for the given combination of insurers and medical providers. The baseline on which these percentages are based is when medical care is not sold at all, and all medical needs shocks are experienced as pain.

Table 4: Individuals' Welfare by Number of Insurers and Providers, Large Medical Needs

	# of Medical Providers								
# of Insurers	1	2	3	4	5	6	7	8	9
0	5.27%	7.91%	8.99%	9.66%	10.11%	10.43%	10.68%	10.87%	11.02%
1	1.22%	9.22%	9.22%	9.22%	9.22%	9.22%	9.22%	9.22%	9.22%
2	3.33%	8.17%	8.80%	9.16%	9.39%	9.54%	9.66%	9.75%	9.81%
3	4.47%	8.37%	9.18%	9.64%	9.94%	10.15%	10.30%	10.42%	10.52%
4	4.97%	8.38%	9.26%	9.77%	10.10%	10.34%	10.51%	10.64%	10.75%
5	5.24%	8.37%	9.29%	9.82%	10.17%	10.42%	10.61%	10.75%	10.86%
6	5.42%	8.36%	9.30%	9.85%	10.21%	10.47%	10.66%	10.81%	10.93%
7	5.54%	8.35%	9.30%	9.87%	10.24%	10.50%	10.70%	10.85%	10.97%
8	5.63%	8.34%	9.30%	9.86%	10.25%	10.52%	10.72%	10.88%	11.00%
9	5.70%	8.33%	9.30%	9.88%	10.26%	10.54%	10.74%	10.89%	11.02%

Table 5: Individuals' Welfare by Number of Insurers and Providers, Large Preference Shocks

	# of Medical Providers								
# of Insurers	1	2	3	4	5	6	7	8	9
0	4.51%	6.75%	7.62%	8.14%	8.49%	8.74%	8.93%	9.08%	9.20%
1	1.64%	7.20%	7.20%	7.20%	7.20%	7.20%	7.20%	7.20%	7.20%
2	0.30%	4.00%	4.55%	4.85%	5.04%	5.17%	5.27%	5.35%	5.41%
3	-0.25%	3.01%	3.63%	3.99%	4.22%	4.38%	4.50%	4.59%	4.67%
4	-0.43%	2.56%	3.21%	3.58%	3.82%	3.99%	4.12%	4.22%	4.30%
5	-0.49%	2.34%	2.99%	3.37%	3.61%	3.79%	3.92%	4.02%	4.10%
6	-0.52%	2.22%	2.86%	3.24%	3.49%	3.67%	3.80%	3.90%	3.99%
7	-0.53%	2.15%	2.79%	3.17%	3.42%	3.60%	3.73%	3.83%	3.91%
8	-0.53%	2.10%	2.74%	3.12%	3.37%	3.55%	3.68%	3.78%	3.87%
9	-0.54%	2.07%	2.71%	3.09%	3.34%	3.51%	3.65%	3.75%	3.83%

Table 6: Individuals' Welfare by Number of Insurers and Providers, Small Preference Shocks

	# of Medical Providers								
# of Insurers	1	2	3	4	5	6	7	8	9
0	4.51%	6.75%	7.62%	8.14%	8.49%	8.74%	8.93%	9.08%	9.20%
1	0.84%	7.11%	7.11%	7.11%	7.11%	7.11%	7.11%	7.11%	7.11%
2	1.71%	5.78%	6.27%	6.55%	6.72%	6.85%	6.94%	7.00%	7.06%
3	1.89%	5.97%	6.59%	6.95%	7.21%	7.34%	7.46%	7.55%	7.62%
4	1.95%	5.99%	6.67%	7.07%	7.32%	7.50%	7.64%	7.74%	7.82%
5	2.15%	6.00%	6.71%	7.12%	7.39%	7.58%	7.73%	7.84%	7.92%
6	2.28%	5.99%	6.72%	7.15%	7.43%	7.63%	7.78%	7.89%	7.98%
7	2.37%	5.99%	6.73%	7.17%	7.46%	7.66%	7.81%	7.93%	8.02%
8	2.44%	5.98%	6.74%	7.18%	7.47%	7.68%	7.83%	7.95%	8.05%
9	2.50%	5.98%	6.74%	7.19%	7.49%	7.70%	7.85%	7.97%	8.07%

Note: Tables 5, 6, and 4 present individuals' welfare under alternative parameter sets: when medical needs shocks are larger than baseline ($\lambda = 0.5$), when preference shocks (over insurance plans) are larger than baseline ($\sigma = 0.1$), and when they are smaller than baseline ($\sigma = 0.025$), respectively. See Section 4.2.

Table 7: Individuals' Welfare by Number of Insurers and Providers, No Bargaining Effect

	# of Medical Providers									
# of Insurers	1	2	3	4	5	6	7	8	9	C
0	4.51%	6.75%	7.62%	8.14%	8.49%	8.74%	8.93%	9.08%	9.20%	N/A%
1	1.08%	3.95%	4.84%	5.38%	5.74%	6.00%	6.19%	6.34%	6.47%	7.56%
2	0.84%	3.32%	4.08%	4.56%	4.89%	5.13%	5.31%	5.45%	5.57%	6.60%
3	0.97%	3.60%	4.28%	4.70%	4.99%	5.21%	5.37%	5.49%	5.59%	6.53%
4	1.02%	3.80%	4.49%	4.91%	5.19%	5.39%	5.54%	5.66%	5.76%	6.65%
5	1.04%	3.93%	4.62%	5.04%	5.33%	5.53%	5.68%	5.80%	5.89%	6.76%
6	1.05%	4.01%	4.71%	5.14%	5.42%	5.62%	5.78%	5.90%	5.99%	6.85%
7	1.05%	4.07%	4.78%	5.21%	5.49%	5.70%	5.85%	5.97%	6.06%	6.92%
8	1.05%	4.11%	4.83%	5.26%	5.54%	5.75%	5.90%	6.02%	6.12%	6.98%
9	1.05%	4.15%	4.87%	5.30%	5.59%	5.79%	5.94%	6.06%	6.16%	7.02%

Table 8: Individuals' Welfare by Number of Insurers and Providers, Only Competitive Effect

	# of Medical Providers								
# of Insurers	1	2	3	4	5	6	7	8	9
0	4.51%	6.75%	7.62%	8.14%	8.49%	8.74%	8.93%	9.08%	9.20%
1	4.46%	6.33%	7.06%	7.50%	7.79%	8.00%	8.16%	8.28%	8.38%
2	5.94%	7.40%	7.95%	8.32%	8.55%	8.72%	8.85%	8.95%	9.03%
3	7.04%	8.30%	8.79%	9.09%	9.29%	9.43%	9.54%	9.63%	9.70%
4	7.70%	8.88%	9.34%	9.61%	9.80%	9.93%	10.03%	10.11%	10.17%
5	8.14%	9.27%	9.71%	9.97%	10.15%	10.28%	10.37%	10.45%	10.51%
6	8.43%	9.55%	9.97%	10.23%	10.40%	10.52%	10.62%	10.69%	10.75%
7	8.65%	9.75%	10.17%	10.42%	10.59%	10.71%	10.80%	10.87%	10.93%
8	8.81%	9.91%	10.32%	10.57%	10.73%	10.85%	10.94%	11.01%	11.07%
9	8.94%	10.03%	10.44%	10.69%	10.85%	10.97%	11.06%	11.13%	11.18%

Table 9: Individuals' Welfare by Number of Insurers and Providers, No Demand Effect

	# of Medical Providers								
# of Insurers	1	2	3	4	5	6	7	8	9
0	4.51%	6.75%	7.62%	8.14%	8.49%	8.74%	8.93%	9.08%	9.20%
1	4.46%	9.27%	9.27%	9.27%	9.27%	9.27%	9.27%	9.27%	9.27%
2	5.94%	8.32%	8.72%	8.95%	9.09%	9.19%	9.27%	9.32%	9.37%
3	7.04%	8.79%	9.20%	9.43%	9.59%	9.70%	9.78%	9.84%	9.89%
4	7.70%	9.21%	9.61%	9.85%	10.00%	10.11%	10.19%	10.16%	10.30%
5	8.14%	9.52%	9.92%	10.15%	10.09%	10.41%	10.49%	10.55%	10.60%
6	8.43%	9.75%	10.14%	10.37%	10.52%	10.63%	10.71%	10.78%	10.77%
7	8.65%	9.92%	10.31%	10.54%	10.69%	10.80%	10.88%	10.94%	10.99%
8	8.81%	10.05%	10.44%	10.67%	10.82%	10.93%	11.01%	11.08%	11.13%
9	8.94%	10.16%	10.55%	10.78%	10.93%	11.04%	11.12%	11.18%	11.23%

Note: Tables 7, 8, and 9 present a decomposition of individuals' welfare as the bargaining effect and demand effect are turned off. The competitive effect is operative in all tables. See Section 4.3.

Table 10: Individuals' Welfare by Number of Insurers and Providers, Observable Needs

	# of Medical Providers								
# of Insurers	1	2	3	4	5	6	7	8	9
0	4.51%	6.75%	7.62%	8.14%	8.49%	8.74%	8.93%	9.08%	9.20%
1	1.68%	10.21%	10.21%	10.21%	10.21%	10.21%	10.21%	10.21%	10.21%
2	3.98%	11.31%	11.87%	12.19%	12.37%	12.51%	12.61%	12.68%	12.74%
3	5.44%	12.37%	13.03%	13.39%	13.63%	13.79%	13.91%	14.00%	14.07%
4	6.18%	12.87%	13.59%	14.00%	14.26%	14.44%	14.58%	14.68%	14.76%
5	6.62%	13.12%	13.91%	14.35%	14.63%	14.82%	14.96%	15.07%	15.16%
6	6.91%	13.29%	14.11%	14.57%	14.86%	15.06%	15.21%	15.33%	15.42%
7	7.12%	13.39%	14.24%	14.72%	15.02%	15.23%	15.38%	15.50%	15.60%
8	7.27%	13.47%	14.34%	14.82%	15.13%	15.35%	15.51%	15.63%	15.73%
9	7.39%	13.53%	14.41%	14.90%	15.22%	15.44%	15.60%	15.73%	15.83%

Table 11: Price of Care by Number of Insurers and Medical Providers, Observable Needs

	# of Medical Providers								
# of Insurers	1	2	3	4	5	6	7	8	9
0	1.607	1.092	1.076	1.066	1.060	1.056	1.052	1.050	1.048
1	1.750	1.008	1.008	1.008	1.008	1.008	1.008	1.008	1.008
2	2.003	1.129	1.101	1.086	1.078	1.071	1.067	1.064	1.061
3	2.162	1.201	1.158	1.135	1.121	1.112	1.105	1.100	1.096
4	2.258	1.241	1.190	1.163	1.147	1.136	1.128	1.121	1.117
5	2.318	1.267	1.211	1.181	1.163	1.150	1.142	1.135	1.129
6	2.365	1.284	1.224	1.193	1.173	1.160	1.151	1.144	1.138
7	2.402	1.296	1.234	1.201	1.181	1.167	1.158	1.150	1.144
8	2.422	1.305	1.241	1.208	1.187	1.173	1.163	1.155	1.149
9	2.439	1.312	1.247	1.212	1.191	1.177	1.167	1.159	1.152

Table 12: Equilibrium Outcomes Under Perfect Competition and Social Planner's Solution

Scenario	p	D	z	c	t	ΔWelfare
Private information, perfect competition	1.461	0.487	0.463	0.510	N/A	11.30%
Private information, social planner	1.330	0.443	0.278	0.703	N/A	12.20%
Observable needs, perfect competition	1.297	0.432	0.560	0	0.210	24.80%
Observable needs, social planner	1.140	0.380	0.434	0	0.242	26.25%

Note: Tables 10 and 11 present individual welfare and equilibrium prices in an alternative model in which individuals' level of medical need η_i is observable by insurers. Insurance contracts are thus written conditional on medical need: care is fully reimbursed, with no copay, when need exceeds the contracted threshold level. See Section 4.4 and Appendix E. Table 12 summarizes outcomes when there is perfect competition (individuals have no preference shocks over insurers) or under the social planner's optimal insurance contract in both the private information (baseline) and observable needs models. See Appendix F.

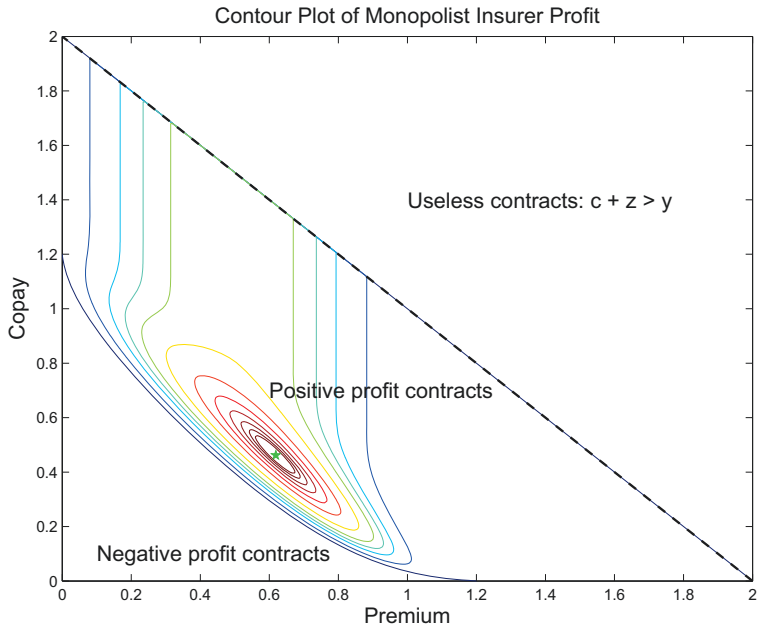

Figure 1: Contour plot of monopolist insurer's expected profit across contracts offered; green star is the profit-maximizing contract. Only non-negative contours are shown. Contracts whose premium and copay sum to exceed income are unusable by individuals.

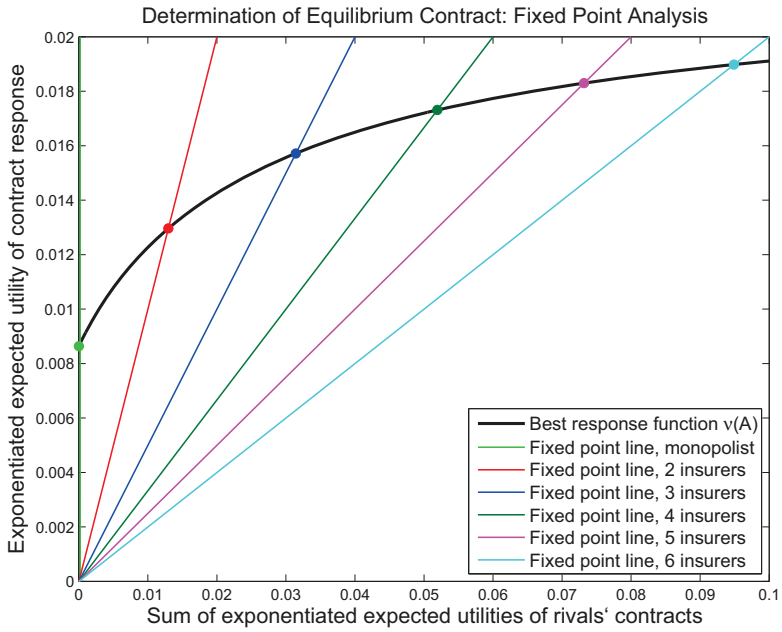

Figure 2: Determination of the equilibrium contract by fixed point analysis. With $N-1$ rival firms, symmetric equilibrium is defined by $\tilde{u} = \nu((N-1)\tilde{u})$, where $\nu(\cdot)$ is the best response function to the sum of exponentiated expected utilities of opponents' contracts.

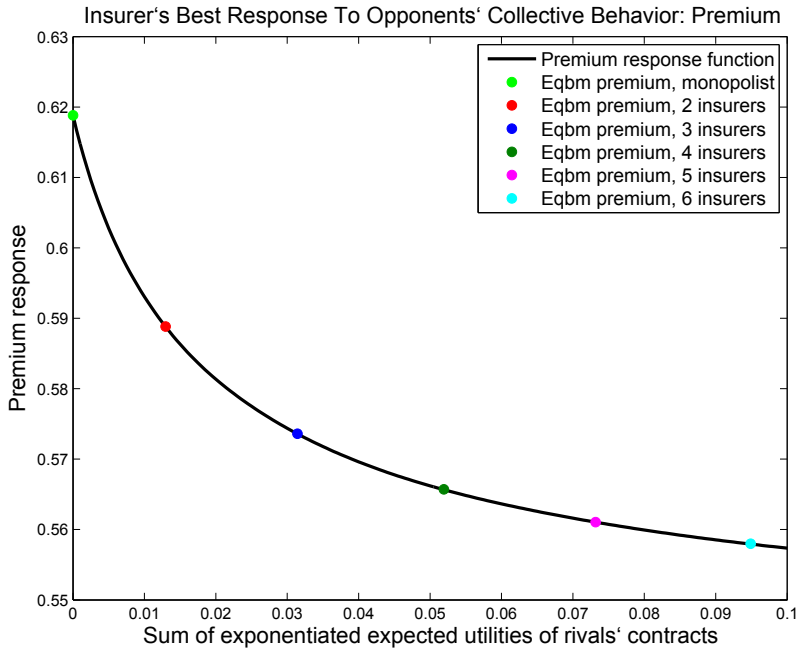

Figure 3: Insurer's best response premium for different levels of sum of exponentiated expected utilities of opponents' contracts, at base parameters. Equilibrium outcomes for different levels of insurers are included (see Figure 2).

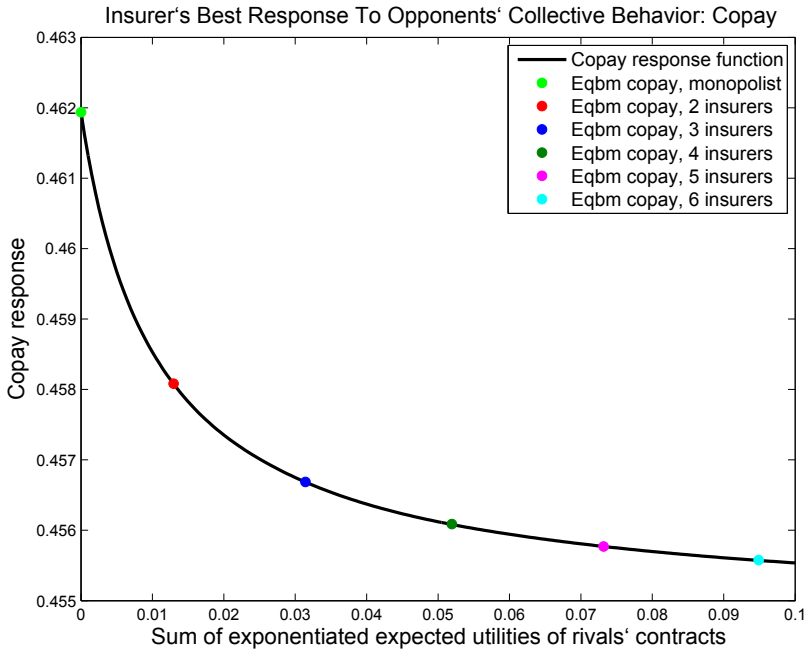

Figure 4: Insurer's best response copay for different levels of sum of exponentiated expected utilities of opponents' contracts, at base parameters. Equilibrium outcomes included.

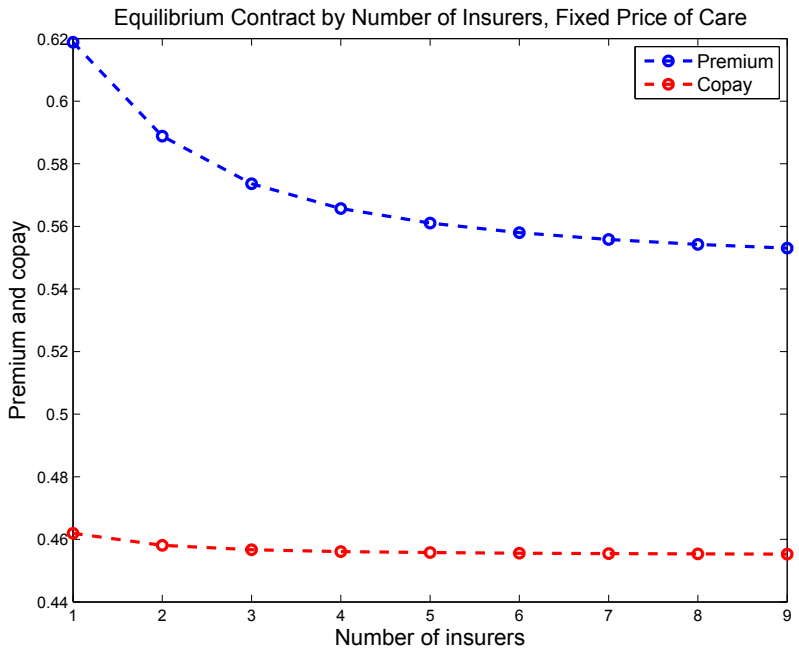

Figure 5: Equilibrium premium and copay by number of insurers at base parameters.

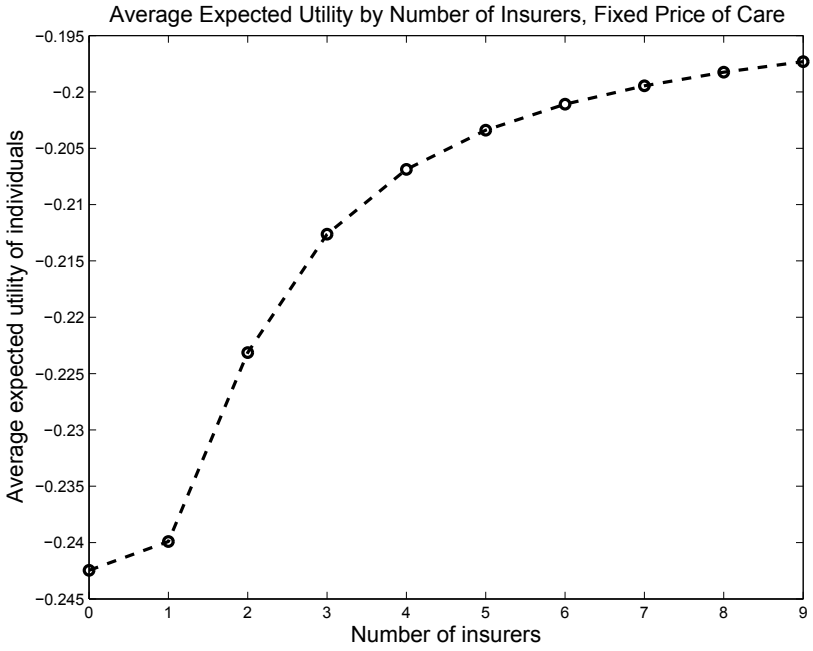

Figure 6: Average expected utility of individuals by number of insurers at base parameters.

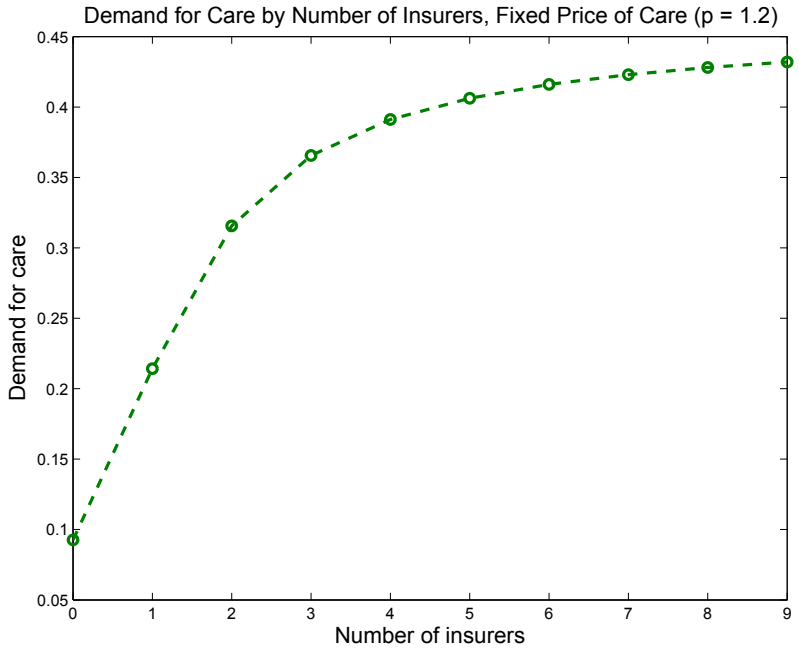

Figure 7: Total demand for medical care by number of insurers at the baseline fixed price ($p = 1.2$). Additional insurers increase demand for care.

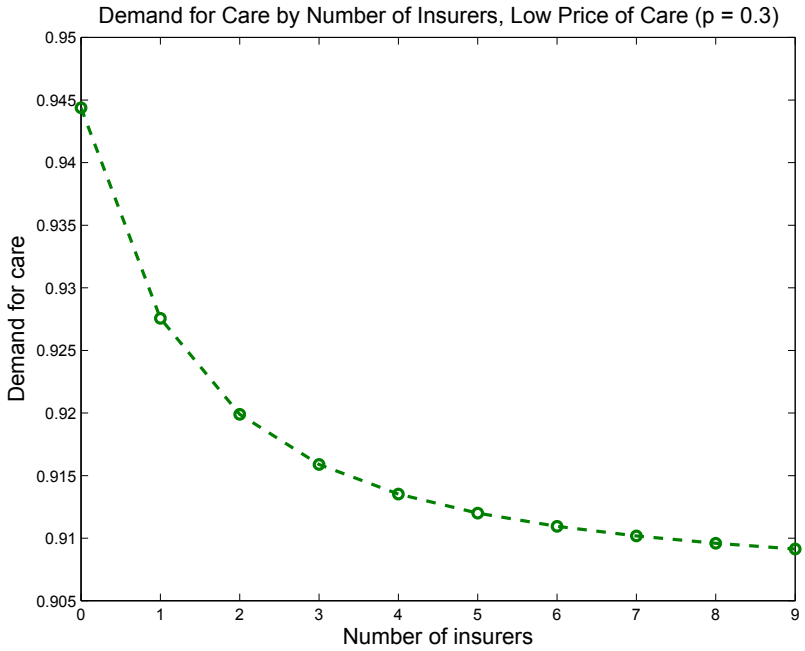

Figure 8: Total demand for medical care by number of insurers at a very low price ($p = 0.3$); additional insurers reduce demand for care.

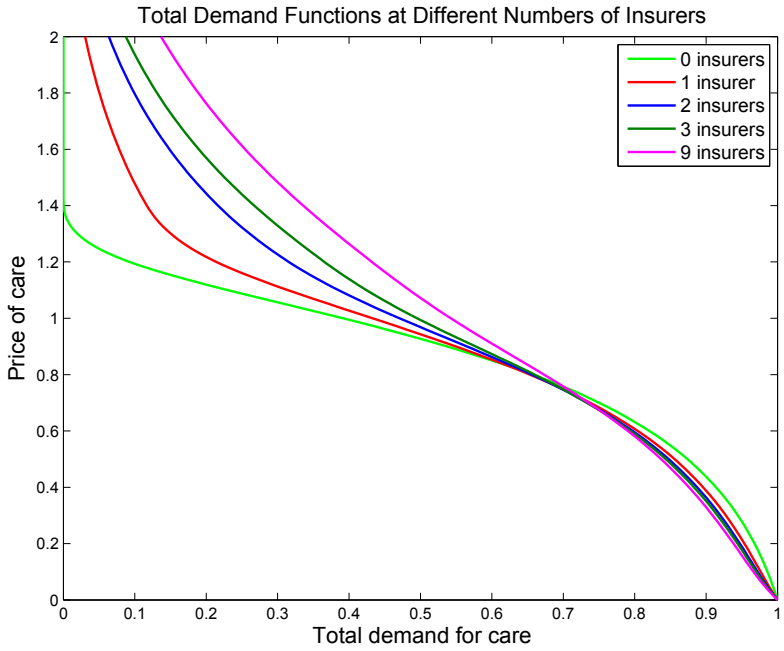

Figure 9: Total demand for care as a function of price, across different numbers of insurers.

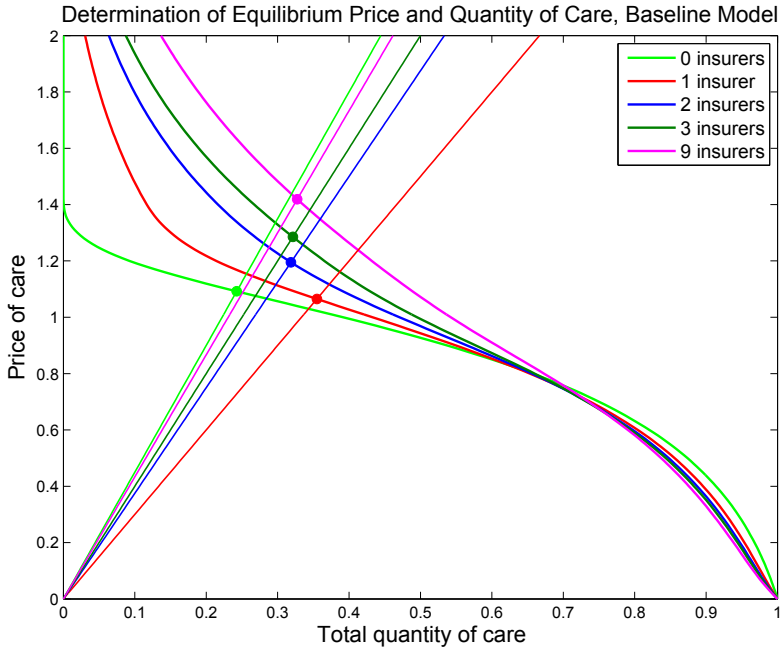

Figure 10: Determination of equilibrium price and quantity of care at different numbers of insurers. Demand curves are as above. Rays from origin represent pseudo-supply curves as in (26): ceiling of equilibrium prices characterized by $D = p/(\gamma(1 + \frac{1}{M} - \frac{1}{MN}))$ when $N > 0$, or $D = p/(\gamma(1 + \frac{1}{M}))$ when there are no insurers. Shown with $M = 2$.

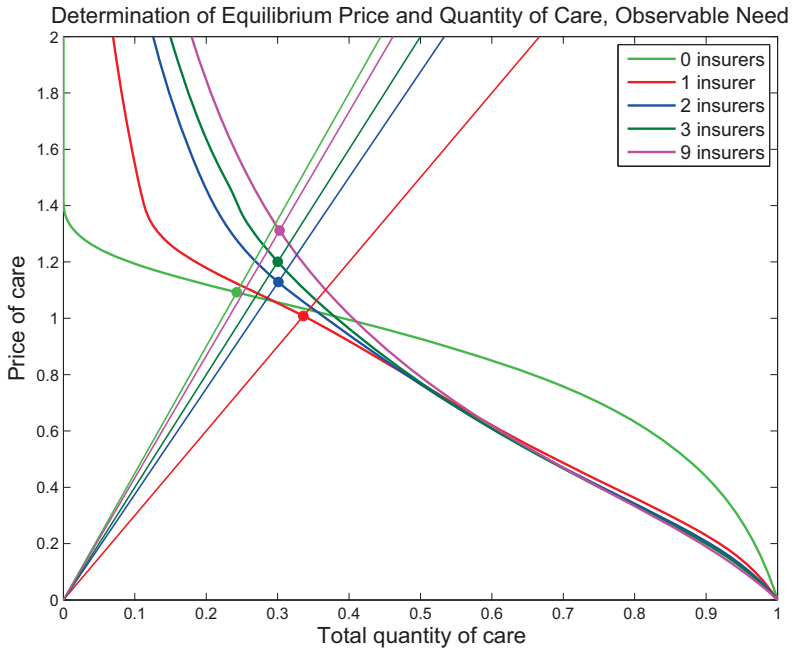

Figure 11: Determination of equilibrium price and quantity of care at different numbers of insurers in the observable needs model. See Section 4.4.

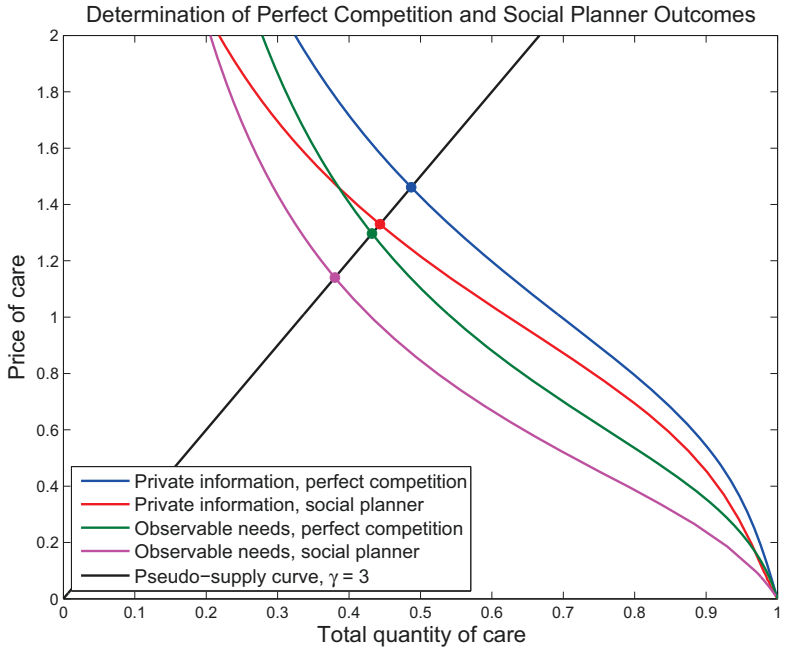

Figure 12: Determination of equilibrium price and quantity of care at different numbers of insurers under perfect competition and the social planner's solution in the baseline model and the observable medical needs specification. Insurers are assumed to collude against non-monopolist providers, and all individuals purchase their contract. See Appendix F.

Appendices

A Properties of the Best Response Contract

In Section 3.2, it was asserted that there exists a non-negative best response contract for any menu of rival insurers' reduced form contracts X_{-j} at any price $\underline{p}$. This appendix provides proof of the existence of this best response, discusses why it is almost certainly unique, and establishes its monotonicity with respect to A.

A.1 Existence of Non-Negative Profit Best Response

If an insurer offered a contract such that $y < z$, then individuals could never purchase care and thus no one would buy the contract, forcing expected profit to zero. Thus if any insurer would earn negative profit, he can instead offer a contract that no one would purchase and improve his payoff to zero. In equilibrium, no insurer will earn negative profit. Moreover, if at least one insurer would earn positive profit while another would earn zero profit, the latter can improve his payoff by duplicating the former's contract. The existence of at least one contract that yields positive profit is easy to establish (as long as $\underline{p} < y$), as the trivial contract $(0, \underline{p})$ returns zero profit, and the isoprofit curve through this contract has finite slope. We can move along this isoprofit curve, and a positive profit contract is guaranteed to be found in any epsilon neighborhood of any point on the curve due to the continuity of the profit function. Thus in equilibrium, all insurers will offer feasible contracts that yield positive profit. In searching for a symmetric equilibrium, we can thus assume an interior solution, necessitating that the first order conditions are satisfied.

Demonstrating the existence of a positive profit best response to any combination of rival insurers' contracts is straightforward. The isoprofit curve corresponding to zero profit (beginning at $(0, \underline{p})$ and ending at $(\underline{p}, 0)$) creates a lower closed boundary, while the $y = c$ and $y = z$ lines serve as an upper closed boundary. As contracts approach $y = z$, their expected profit falls to zero as fewer and fewer individuals purchase that contract (because $\overline{u}(\chi) \to -\infty$ as $z \to y$). When the premium and copay are so high that $y < z + c$, changes in the copay have no effect on $\overline{u}$ or profit because no customers were buying care anyway. Indeed, these "useless contracts" would almost never be purchased by individuals; only those with an extremely large preference shock would purchase a contract that prevents them from buying care. If insurers are barred from offering a negative premium or copay,[16]

[16] Even if this were allowed, it could never occur in equilibrium. Risk averse individuals prefer insurance to anti-insurance, so this contract can always be trumped.

then this creates a compact set of non-negative profit contracts; with a continuous expected profit function that has at least one positive profit contract, we are guaranteed to find a profit-maximizing contract on the interior of this set, at which the first order conditions will hold.

A.2 First and Second Order Conditions of Profit Function

Using the definition (as given in (13)) of insurer j's expected profit when offering contract χ_j and rivals' offered menu of contracts is X_{-j}, the first order conditions for an optimal response to competitors' offered contracts are:

$$\frac{\partial \pi(\chi_j|X_{-j})}{\partial z_j} = \frac{\partial r(\chi_j)}{\partial z_j} q(\chi_j|X_{-j}) + \frac{\partial q(\chi_j|X_{-j})}{\partial z_j} r(\chi_j) = 0, \tag{27}$$

$$\frac{\partial \pi(\chi_j|X_{-j})}{\partial c_j} = \frac{\partial r(\chi_j)}{\partial c_j} q(\chi_j|X_{-j}) + \frac{\partial q(\chi_j|X_{-j})}{\partial c_j} r(\chi_j) = 0. \tag{28}$$

To establish the uniqueness of a solution to the first order conditions (and thus that the best response to any reasonable offerings from competitors is a singleton), a sufficient condition is that the second order conditions hold everywhere– that the profit function is strictly concave everywhere on the interior of the set of admissible contracts. These second order conditions are:

$$\frac{\partial^2 \pi(\chi_j|X)}{\partial z_j^2} = \frac{\partial^2 r(\chi_j)}{\partial z_j^2} q(\chi_j|X) + 2\frac{\partial r(\chi_j)}{\partial z_j}\frac{\partial q(\chi_j|X)}{\partial z_j} + \frac{\partial^2 q(\chi_j|X)}{\partial z_j^2} r(\chi_j) < 0, \tag{29}$$

$$\frac{\partial^2 \pi(\chi_j|X)}{\partial c_j^2} = \frac{\partial^2 r(\chi_j)}{\partial c_j^2} q(\chi_j|X) + 2\frac{\partial r(\chi_j)}{\partial c_j}\frac{\partial q(\chi_j|X)}{\partial c_j} + \frac{\partial^2 q(\chi_j|X)}{\partial c_j^2} r(\chi_j) < 0, \tag{30}$$

$$\frac{\partial^2 \pi(\chi_j|X)}{\partial z_j^2}\frac{\partial^2 \pi(\chi_j|X)}{\partial c_j^2} - \left(\frac{\partial^2 \pi(\chi_j|X)}{\partial z_j \partial c_j}\right)^2 < 0. \tag{31}$$

The various first and second derivatives of precursor functions referenced in the first and second order conditions are as follows, with analysis of their sign (arguments and subscripts are suppressed for brevity and clarity):

$$\frac{\partial r}{\partial z} = 1 + \frac{\partial \Delta u}{\partial z} f(\Delta u)(\underline{p} - c) > 0, \qquad \frac{\partial r}{\partial c} = (1 - F(\Delta u)) + \frac{\partial \Delta u}{\partial c} f(\Delta u)(\underline{p} - c) > 0. \tag{32}$$

$$\frac{\partial^2 r}{\partial z^2} = \left(\underbrace{\frac{\partial^2 \Delta u}{\partial z^2} f(\Delta u)}_{+} + \underbrace{\frac{\partial \Delta u}{\partial z} f'(\Delta u)}_{-\text{ if } f \text{ is Weibull}} \right) (\underline{p} - c), \tag{33}$$

$$\frac{\partial^2 r}{\partial c^2} = \left(\underbrace{\frac{\partial^2 \Delta u}{\partial c^2} f(\Delta u)}_{+} + \underbrace{\frac{\partial \Delta u}{\partial c} f'(\Delta u)}_{-\text{ if } f \text{ is Weibull}} \right) (\underline{p} - c) - \underbrace{\left(\frac{\partial \Delta u}{\partial c} + 1 \right) f(\Delta u)}_{+}. \tag{34}$$

$$\frac{\partial q}{\partial z} = \frac{\partial \overline{u}}{\partial z}(q - q^2) < 0, \qquad \frac{\partial q}{\partial c} = \frac{\partial \overline{u}}{\partial c}(q - q^2) < 0. \tag{35}$$

$$\frac{\partial^2 q}{\partial z^2} = \frac{\partial^2 \overline{u}}{\partial z^2} \underbrace{(q - q^2)}_{+} + \underbrace{\frac{\partial \overline{u}}{\partial z}(2q^3 - 3q^2 + q)}_{\pm \text{ as } q \gtrless \frac{1}{2}}, \qquad \frac{\partial^2 q}{\partial c^2} = \frac{\partial^2 \overline{u}}{\partial c^2}(q - q^2) + \frac{\partial \overline{u}}{\partial c}(2q^3 - 3q^2 + q). \tag{36}$$

$$\frac{\partial \Delta u}{\partial z} = u_1' - u_0' > 0, \qquad \frac{\partial \Delta u}{\partial c} = u_1' > 0, \qquad \frac{\partial^2 \Delta u}{\partial z^2} = \underbrace{u_0'' - u_1'' > 0}_{\text{if } u \text{ is CRRA}}, \qquad \frac{\partial^2 \Delta u}{\partial c^2} = -u_1'' > 0. \tag{37}$$

$$\frac{\partial \overline{u}}{\partial z} = -\left(F(\Delta u))u_0' + (1 - F(\Delta u))u_1'\right) < 0, \qquad \frac{\partial \overline{u}}{\partial c} = -(1 - F(\Delta u))u_1' < 0. \tag{38}$$

$$\frac{\partial^2 \overline{u}}{\partial z^2} = \underbrace{F(\Delta u))u_0'' + (1 - F(\Delta u))u_1''}_{-} + \underbrace{\left(\frac{\partial \Delta u}{\partial z} \right)^2 f(\Delta u)}_{+}, \tag{39}$$

$$\frac{\partial^2 \overline{u}}{\partial c^2} = \underbrace{(1 - F(\Delta u))u_1''}_{-} + \underbrace{\left(\frac{\partial \Delta u}{\partial c} \right)^2 f(\Delta u)}_{+}. \tag{40}$$

The parameterization of $f(\cdot)$ used in Section 4 is the exponential distribution, a special case of the Weibull distribution when the shape parameter is set to 1. Thus the signs of some terms above are shown when $f(\cdot)$ is distributed as a Weibull.

A.3 Uniqueness of Best Response

Rather than conduct a long and fruitless mathematical analysis, it is worth pointing out that the profit function *cannot* be concave everywhere. For a "useless" contract such that $y < z + c$, expected profit per customer is certainly positive (as premiums are being paid, but customers never buy care), and the preference shock for contracts guarantees that the share of customers will be positive but tiny, approaching zero as the premium increases. That is, $\pi(\chi|X)$ will be strictly positive when

$y < z + c$; this requires that $\frac{\partial^2 \pi}{\partial z^2} > 0$ at some point, violating the concavity of expected profit. Instead of concavity, an alternate sufficient condition must be demonstrated. Possibilities include showing that $\pi(\chi|X)$ is strictly quasi-concave, or that the first order conditions (27) and (28) have a single-crossing property. Either of these will guarantee that the best response to any set of contracts offered by rival insurers is a singleton.

No formal proofs of these conditions are available at this time and may not exist,[17] but a contour map of a monopolist insurer's profit function is presented in Figure 1. As insurers will never sell a contract that yields negative expected profit, only non-negative contour lines are shown. The zero isoprofit curve is the lowest shown, spanning from $(\underline{p}, 0)$ to $(0, \underline{p})$ as described above. Contour lines are also omitted for "useless contracts", when the premium and copay sum to exceed income so that individuals who purchase them can never afford care. As noted above, these contracts would be purchased by nearly no individuals (only those unlucky few with very extreme preference shocks) and thus yield near-zero profit. The contours of Figure 1 demonstrate that the profit function is not quasi-concave, as there are some upper contour sets that are non-convex. However, the function is single peaked and is quasi-concave on the subset of contracts that are beneficial to individuals by offering at least some risk protection: contracts such that $z + c < \underline{p}$. That is, the violations of quasi-concavity occur quite far away from the peak. These patterns are not specific to the monopolist case (nor to the particular parameter set used), and are consistent when rival firms offer competing contracts. The location of the peak shifts inward to more consumer-favorable contracts, but the profit function continues to be single peaked and quasi-concave among reasonable contracts.

A.4 Monotonicity of Best Response Function

The components of the unique contract that is a best response to a particular menu of rivals' contracts X_{-j} is defined by:

$$(\check{z}(A), \check{c}(A)) = \arg\max_{z,c} \pi((z,c)|X_{-j}), \qquad A = \sum_{\ell \neq j,0} \exp(\overline{u}(\chi_\ell)). \tag{41}$$

To determine the shape of these functions, consider a contract on it, representing the best response contract when the sum of rivals' exponentiated expected utilities is equal to some value $A \geq 0$. Thus

[17] Caplin and Nalebuff (1991) presents a proof of a pure strategy equilibrium in a market with spatially differentiated products, even in the presence of consumer heterogeneity. However, they treat these product characteristics as exogenous rather than strategic choices. Thus their proof applies in a situation where copayments are fixed and premia are the only strategic choice variable; the extension in my model to a multi-dimensional strategic space complicates the analysis. Moreover, no proof of uniqueness is offered.

the first order conditions for an optimal best response hold at this contract. Substituting (35) into (28) and (27), we have:

$$\frac{\partial r_j}{\partial c_j} q_j + \frac{\partial \overline{u}_j}{\partial c_j}(q_j - q_j^2) r_j = 0. \tag{42}$$

$$\frac{\partial r_j}{\partial z_j} q_j + \frac{\partial \overline{u}_j}{\partial z_j}(q_j - q_j^2) r_j = 0. \tag{43}$$

The first term of the LHS of both equations is positive, while the second term is of equal magnitude but negative. If A is increased very slightly to A', how does the best response contract change? When rivals' contracts offer better expected utility for individuals, insurer j's share of individuals q will decrease, while the profit per customer r_j and the partial derivatives in the equations will be unchanged. But when q_j decreases due to rivals' slightly better contracts, $(q_j - q_j^2)$ decreases by less, so that the second term has greater magnitude. Thus when A increases to A', the LHS of (42) and (43) becomes negative, so that insurer j's expected profit is increasing as both the premium and copay are reduced. Because $\pi(\chi|X_{-j})$ is locally quasi-concave, the best response contract to A' must have lower premium and copay than the best response to A. If both the premium and the copay of the new contract are lower, an individual purchasing the new contract attains higher expected utility. In summary, the locus of best response contracts is strictly monotone, and the expected utilities offered to individuals by insurers through their contracts are strategically complementary.

B Properties of Symmetric Equilibrium for Insurers

Section 3.2 demonstrates that there can be a symmetric equilibrium among insurers in which each one offers the same contract. This appendix establishes several properties of the insurers' symmetric equilibrium: that there are only symmetric equilibria, that a symmetric equilibrium exists, and that it is very likely unique.

B.1 Impossibility of Non-Symmetric Equilibria

To show that only symmetric equilibria are possible, consider an alleged equilibrium with two insurers selling different reduced form contracts. Because it is an equilibrium, each insurer must be offering the best response contract to his rivals' choices. Suppose insurer 1's contract offers higher expected utility to individuals than insurer 2's contract. Then the sum of exponentiated expected utilities of insurer 1's rivals' contracts must be lower than that of insurer 2 (as the sets of rivals only differ

by these two insurers). As the previous appendix showed that expected utilities of contracts are strategic complements, this implies that insurer 1's best response contract gives lower expected utility than insurer 2's, contradicting the supposition that insurer 1 gives higher expected utility (or that these are both best responses to the other). Thus any two insurers must offer the same expected utility to individuals in equilibrium, and indeed identical contracts.

B.2 Existence of a Symmetric Equilibrium

To establish that a symmetric equilibrium exists, there must be at least one solution to (14). The proof will be performed by applying the intermediate value theorem. When rivals offer only contracts with $\overline{u} = -\infty \implies \exp(\overline{u}) = \tilde{u} = 0$, this is equivalent to when there is only a single monopolist insurer who will offer a contract with finite exponentiated utility $\nu_{\underline{p}}(0) > 0$. This utility must be positive, as it was established above that there exists a non-negative profit contract in response to any menu, and thus individuals can attain expected utility above $-\infty$. In the presence of competition, the symmetric equilibrium is defined by a fixed point of a function with positive slope. As $\tilde{u} \to \infty$, the exponentiated utility of the best response contract does not also become arbitrarily large. Rather, the utility of any offered contract is bounded by the zero profit locus. Because the first term of (27) and (28) are always negative, the second term must be positive, with the insurer earning strictly positive profit per customer. Thus the expected utility of a best response contract is bounded above by the highest expected utility contract that yields zero profit. Simple application of the intermediate value theorem requires that there exists at least one solution to (14), as the RHS exceeds the LHS when $\tilde{u} = 0$ but is overtaken at values of $\tilde{u}$ greater than the maximum exponential utility zero profit contract.

B.3 Uniqueness of Symmetric Equilibrium

To ensure that there is only one symmetric equilibrium, a sufficient condition is that $\nu_{\underline{p}}(A)$ is strictly concave. While a formal proof of concavity is not available, an informal semi-proof demonstrates that this property almost certainly holds. Note that as A becomes large, $\nu_{\underline{p}}(A)$ asymptotes to some finite level, thus it must *eventually* be strictly concave. Similarly, the best response function must converge to some contract, so the best response premium and copay functions are eventually concave in A, and possibly always so. Even if these functions are not concave everywhere, (39) and (40) indicate that expected utility is likely increasing concavely as the premium and copay fall when A

rises. Thus even if the best response contract were to change convexly with A for some range, it may be tempered by the concave translation to expected utility. In this way, the function $\nu_{\underline{p}}(A)$ is almost certainly concave so that there is only one symmetric equilibrium. Indeed, it is concave at every price and parameterization tested.

B.4 Other Properties of Insurers' Equilibrium

It is reasonable to expect that when the number of insurers becomes very large, competition causes the equilibrium contract to converge to some ideal point for individuals: the expected utility-maximizing contract on the zero profit locus. However, this is not the case. As N becomes very large, the share of individuals of any insurer approaches zero. Substituting (35) into (29) and (30), taking the limit as $q \to 0$, applying l'Hôpital's rule and rearranging reveals that per customer profit converges to a positive value. Using the version with the premium:

$$\lim_{q \to 0} r(\chi^*) = -\left(\frac{\partial \Delta u}{\partial z}\right)^{-1} - f(\Delta u)(\underline{p} - c) > 0. \tag{44}$$

A similar equation using the copay can also be found. No matter how many competitors there are, the equilibrium contract will always yield positive per customer profits.

If the magnitude of the utility function (and the medical need shocks) is increased relative to the size of the preference shocks, so that differences in expected utility between contracts become greater, then individuals are effectively "better shoppers", less likely to purchase suboptimal contracts. That is, when the preference shocks are small relative to the utility function, competition among many insurers will result in an equilibrium contract with very small per customer profit– the preference shocks act as a barrier to a perfectly competitive solution. As discussed in Section 2.6, the magnitude of the preference shocks could represent the extent of individuals' imperfect ability to learn about, understand, or have access to the different insurance contracts offered. In Section 4.2, the effects of these barriers to perfect competition on individuals' welfare are explored more fully.

C Properties of the Demand Function

Section 3.2.3 introduced the total demand function $D(\underline{p}, N)$ and noted that it is downward sloping in price but has a more complex relationship with the number of insurers. Proofs and evidence of

these properties are provided in this appendix.

C.1 Demand Function is Negatively Sloped

A critical property of $D(\underline{p}, N)$ is that it is decreasing in price; this will become useful in coming analyses. When the price paid to the medical provider increases slightly, per customer profit decreases (due to higher cost sharing) but the share of individuals buying a contract increases (as the null insurance option is less attractive). Starting from the equilibrium contract when the lowest price is $\underline{p}$, suppose price is raised slightly to $\underline{p}'$. Define $\mathring{u}_1'$ as the marginal utility of consumption under the null contract when care is purchased, with related objects defined likewise. To determine how an insurer should change his contract in response, take the derivative of the first order condition for optimal premium (27):

$$\frac{\partial r}{\partial \underline{p}} = -(1 - F(\Delta u)), \qquad \frac{\partial q}{\partial \underline{p}} = \frac{F(\Delta u_0)\mathring{u}_0' + (1 - F(\Delta u_0))\mathring{u}_1'}{N\tilde{u} + \tilde{u}_0} q = Bq > 0, \tag{45}$$

$$\frac{\partial^2 r}{\partial z \partial \underline{p}} = \frac{\partial \Delta u}{\partial z} f(\Delta u), \qquad \frac{\partial^2 q}{\partial z \partial \underline{p}} = \frac{\partial \overline{u}}{\partial z} Bq(1-2q) = B\frac{\partial \overline{u}}{\partial z}(q - 2q^2) = B\frac{\partial q}{\partial z} - B\frac{\partial \overline{u}}{\partial z} q^2. \tag{46}$$

$$\begin{aligned} \frac{\partial^2 \pi}{\partial z \partial \underline{p}} &= \frac{\partial \Delta u}{\partial z} f(\Delta u) q + \frac{\partial r}{\partial z} Bq + \left(B\frac{\partial q}{\partial z} - B\frac{\partial \overline{u}}{\partial z} q^2 \right) r - \frac{\partial q}{\partial z}(1 - F(\Delta u)) = \\ &= \underbrace{\frac{\partial \Delta u}{\partial z} f(\Delta u) q}_{+} + \underbrace{B\left(\frac{\partial r}{\partial z} q + \frac{\partial q}{\partial z} r \right)}_{0} - \underbrace{B\frac{\partial \overline{u}}{\partial z} q^2 r}_{-} - \underbrace{\frac{\partial q}{\partial z}(1 - F(\Delta u))}_{-} > 0. \end{aligned} \tag{47}$$

The second term of this equation is zero because it is exactly the first order condition for the optimal premium, and the remaining terms all contribute positively. A parallel equation can be derived for the copay when $\underline{p}$ changes, which is also strictly positive. Thus an insurer will want to raise both the premium and copay in response to an increase in $\underline{p}$, holding his rivals' contracts fixed. The best response contract function has a unique fixed point, and the expected utility offered by a contract is strategically complementary, so iterated application of the best response contract function (until convergence at the equilibrium) will only increase the premium and copay further.

In this way, an increase in $\underline{p}$ results in an equilibrium contract that offers lower expected utility to individuals through higher z and c. Moreover, both individuals who purchase insurance and those who do not will be less likely to purchase care after price is raised. The only way that total demand for care $D(\underline{p}, N)$ could increase with $\underline{p}$ is if the price increase induced so many individuals to buy

insurance who previously did not that the drop in propensity to buy care was overwhelmed by the switching behavior. A proof that this cannot occur is not available at this time, but this does not seem to occur in practice.

C.2 Demand Function and the Number of Insurers

Total demand for care is increasing in the number of insurers ($D(\underline{p}, N+1) > D(\underline{p}, N)$) so long as the equilibrium contract induces individuals who hold it to purchase care more often than the uninsured (when $\Delta u_j < \Delta u_0$). To see this, consider the fixed point defined by (14). When N is increased by 1, the RHS is "pulled inward" so that it is higher at any given level of $\tilde{u} > 0$ than before. Thus the equilibrium will occur at a higher level of $\tilde{u}$, with associated lower premium and copay. With greater expected utility from the equilibrium contract (and more insurers offering it), a larger portion of individuals will be (non-null) insured. Moreover, the probability of purchasing care conditional on the contract ($1 - F(\Delta u)$) is decreasing in both z and c, as (37) says that both $\frac{\partial \Delta u}{\partial z}$ and $\frac{\partial \Delta u}{\partial c}$ are positive, and so individuals who purchase the equilibrium contract are more likely to buy care when the number of insurers goes up. Both factors of (16) increase with N, thus $D(\underline{p}, N)$ increases with N. This common situation is shown in Figure 7.

This logic breaks down in the odd situation where individuals purchasing the equilibrium contract are less likely to buy care than the uninsured, which can occur at very low prices, as shown in Figure 8 and the lower range of prices in Figure 9. When the price is very low, even uninsured individuals will almost always purchase care, so no insurance contract can offer much risk protection. Moreover, bad contracts that strictly hurt individuals (because $z + c > \underline{p}$) aren't much worse than being uninsured in absolute terms. Thus in equilibrium insurers will offer bad contracts that make individuals who buy them worse off (after the preference shock) and less likely to purchase care. Each additional insurer offering a bad contract increases the probability that any given individual will receive a preference shock large enough to "trick" them into buying that contract, and thus total demand for care is decreasing in N at very low prices.

D Only Symmetric Equilibria Among Providers

According to the equilibrium strategies from Sections 3.1 and 3.2, only medical providers who offer the lowest price among the competitors will receive any customers, whether insured or not; demand

will be split evenly among them as individuals randomize. Thus if some providers are earning positive profits by charging p_0 while another provider is earning no profit with $p_1 > p_0$, then the latter provider can improve his outcome by choosing p_0 instead. To illustrate, suppose $\hat{M}$ providers are offering the lowest price p_0. They earn profits of:

$$\frac{D(p_0)}{\hat{M}} p_0 - \kappa\left(\frac{D(p_0)}{\hat{M}}\right) > 0. \tag{48}$$

If the provider offering p_1 switches to p_0, it will get:

$$\frac{D(p_0)}{\hat{M}+1} p_0 - \kappa\left(\frac{D(p_0)}{\hat{M}+1}\right) > 0. \tag{49}$$

This expression is greater than zero because the per unit profit is higher when the outsider joins them at p_0: sale price is the same, but the average production cost is lower because $\kappa'(\cdot) > 0$. Each provider might earn less when another provider adopts the lowest price, but it can never force positive profits to become negative. If a provider is earning negative profit, he can improve his payoff by choosing a price higher than the current lowest price to earn zero instead.

It is also possible to rule out as possible equilibria situations in which some providers are earning zero profit when offering lowest price p_0, and another insurer is earning zero profit with $p_1 > p_0$. If this were to happen, the insurer offering p_1 could switch to p_0 and gain positive profits (and boost his competitors into the black as well). To see this, note that per unit profit among the $\hat{M}$ offering p_0 is zero, so that average production cost must exactly equal the price. Thus when the outsider joins the group offering the low price, average production cost will fall below the sale price, resulting in positive per unit profits for all. Thus only all providers offering the same price can possibly be an equilibrium– the equilibrium is guaranteed to be symmetric. This holds whether the simple or threatening strategies are employed by insurers.

E Alternative Model: Observable Medical Needs

Section 4.4 presented an alternative specification in which each individual's medical needs η_i were observable and contractible by insurers. Welfare and price results were presented without elaboration or detail about this model. This appendix fleshes out the alternative specification and its solution.

E.1 Alternative Model Formalized

The formal description of the alternative model is mostly identical to that presented in Section 2; specific differences are laid out here. First, the medical need or pain shock η_i for each individual is not private information, but instead is observable to insurers and insurance contracts can be conditioned on it. Next, the space of permissible contracts for insurers to offer is expanded from $\mathbb{R}_+^{M+1}$ to $\mathbb{R}_+^{M+2}$ to account for the threshold t. A contract is specified as $\hat{\chi}_j = (z, \vec{c}, t)$ and the null contract is $\hat{\chi}_0 = (0, \vec{p}, 0)$. A strategy for an insurer is thus a function $\phi : \mathbb{R}_+^M \longrightarrow \mathbb{R}_+^{M+2}$ that takes a vector of prices for each medical provider and returns a valid contract. Individual i who holds contract $\hat{\chi}_j$ and purchases care from provider k will pay c_{jk} if $\eta_i > t_j$ and p_k otherwise. The expected utility[18] of holding contract $\hat{\chi}_j$ with lowest copay c_j can be expressed as:

$$F(\tau_j)u(y-z_j)+(1-F(\tau_j))u(y-z_j-c_j)-\int_0^{\tau_j}\eta F(\eta)d\eta, \qquad \tau_j = \max(t_j, u(y-z_j)-u(y-z_j-c_j)). \quad (50)$$

Insurer j's expected profit per customer is slightly changed from the original model, as described in Appendix E.2. All aspects of the model from the perspective of medical providers are identical to the original model.

E.2 Equilibrium Insurance Contracts Are Complete

In the baseline model, insurers could not profitably offer useful contracts with no copay. If individuals pay nothing out of pocket for care, they will always purchase it regardless of how small their medical needs shock η_i is; this contract is only profitable if the premium is set at or above $\underline{p}$. Moral hazard from lowering the out-of-pocket price of medical care makes complete insurance effectively impossible. In contrast, when η_i is observable and contractible, (reduced form) insurance contracts are written as a triplet of the premium, copay, and threshold level of pain: $\chi = (z, c, t)$. When $\eta_i < t$, the insurer will not pay for care and the individual must pay the full price. While this specification may seem more complex due to the addition of a third contract variable, it can be shown that all copays will be zero in equilibrium, and so reduced form contracts are only an ordered pair (z, t).

Consider the analogue to (12) for this model, expected per customer profit of contract χ_j:

$$r(\chi_j) = z_j - (1 - F(\max(\Delta u_j, t_j)))(\underline{p} - c_j). \quad (51)$$

[18]This equation ignores a corner case in which t_j is so low that individuals choose to sometimes pay the full price of care out of pocket. It does not arise in equilibrium for the model simulated in Section 2.6.

Rather than paying for care $(1 - F(\Delta u_j))$ portion of the time, the insurer only pays when the threshold level of medical need is exceeded. The threshold only binds when $t_j > \Delta u_j$, otherwise this equation is identical to (12). Suppose this condition is met at contract χ_j, and the insurer considers changing the premium and copay so as to leave per customer profit unchanged (while holding the threshold fixed). Using the implicit function theorem, the rate at which the copay should be changed relative to the premium is:

$$\left.\frac{\mathrm{d}c}{\mathrm{d}z}\right|_{\Delta r(\chi)=0} = -\frac{1}{1-F(t)}. \tag{52}$$

When the contract changes in this way, individuals purchasing it see their marginal utility change at a rate of:

$$\left.\frac{\mathrm{d}\overline{u}}{\mathrm{d}z}\right|_{\Delta r(\chi)=0} = -F(t)u_0' + (1-F(t))\left(\frac{1}{1-F(t)} - 1\right)u_1' = F(t)(u_1' - u_0') > 0. \tag{53}$$

Individuals' expected utility improves when the contract is adjusted in this way, as they are shifting consumption from a lower marginal utility state to a higher marginal utility state at the actuarially fair rate.

This property holds for any contract, and is only bounded by $c \geq 0$ both due to restrictions on the space of allowable policies and the fact that the inequality above would flip if a negative copay were reached in this way. If the contract offers higher expected utility and thus a higher share of customers while holding per enrollee profit fixed, expected profit is higher when the contract is changed in this way. From the insurer's perspective, all contracts with a positive copay are dominated by another contract with a lower copay. Thus in equilibrium, only complete contracts with $c = 0$ will be offered. Intuitively, the alternative model works like a simple insurance model in which the probability of loss does not depend on whether the individual is insured, as holding t constant directly controls this probability. Individuals thus prefer more complete insurance when the additional insurance is provided in an actuarially fair way.

E.3 Insurer's Equilibrium In Alternative Model

With the introduction of the threshold t and the above result that the lowest copay of all contracts will be zero in equilibrium, t effectively replaced c as the second policy dimension of reduced form contracts in the alternative model. With this change, similar arguments as those in Section 3.2 can be used to establish that there is a unique equilibrium reduced form contract $\chi^*(\underline{p}) = (z^*(\underline{p}), t^*(\underline{p}))$

for each level of $\underline{p} = \min(\vec{p})$ that medical providers can offer. Analogously to the original model, both the equilibrium premium and threshold are decreasing in the sum of the exponentiated expected utilities of rivals' contracts, and thus both contract dimensions are decreasing in the number of insurers.

The formal equilibrium strategy in the alternative model modifies (15) as follows:

$$\phi(\vec{p}) = (z^*(\underline{p}), \mathbf{1}(\vec{p} > \underline{p}), t^*(\underline{p})), \qquad \underline{p} = \min(\vec{p}). \tag{54}$$

That is, providers who offer the lowest price are assigned a copay of zero, while those choose a higher price are assigned an arbitrarily higher copay that will not be used by any individuals. The premium and threshold follow the optimal policy functions that can be derived using the same logic as Section 3.2. When the threatening equilibrium of Section 3.3.2 is used instead, the strategy is modified from (22) as follows:

$$\tilde{\phi}(\vec{p}) = \begin{cases} (z^*(\underline{p}), 1 - e_M(\text{Rand}(\{1, \cdots, M\})), t^*(\underline{p})) & \text{if } \overline{p} = \underline{p} > p^{**} \\ (z^*(\underline{p}), \mathbf{1}(\vec{p} > \underline{p}), t^*(\underline{p})) & \text{otherwise} \end{cases}, \qquad \overline{p} = \max(\vec{p}). \tag{55}$$

As in the original model, p^{**} is defined by the unique equilibrium price ceiling in (25). All other aspects of the equilibrium for insurers are identical to the original model; welfare analysis proceeds as in Section 4.

F Socially Efficient Outcomes

In Sections 4.2 and 4.3, I presented equilibrium results for four "socially efficient" outcomes: perfect competition and the social planner's solution under both the private information (baseline) and observable medical needs models. This appendix presents derivations of the demand curves described in those sections and shown in Figure 12.

F.1 Private Information and Perfect Competition

Under perfect competition, individuals have infinitesimal preference shocks over insurers' contracts. So long as one contract offers higher utility than all others, that contract will be purchased by all consumers. Insurers are still limited by the zero profit condition and will not offer a contract that

loses money in expectation. As reductions in both z and c are welfare-improving for individuals and profit-reducing for insurers, firms will compete down to the zero profit locus and then along it to the expected utility-maximizing contract on it when market imperfections are removed. To find this contract at any given price of care $\underline{p}$, the following algorithm call be followed.

First, recall the definition of demand for care:

$$D = 1 - F(u(y - z) - u(y - z - c)). \tag{56}$$

In combination with the zero profit condition that says that the premium must equal spending on care, $z = (p - c)D$, this yields:

$$0 = (p - c)(1 - F(u(y - z) - u(y - z - c))) - z. \tag{57}$$

For any given value of c, this equation has only one root in z: the zero profit premium of buying a contract with a given level of generosity (holding $\underline{p}$ fixed). Individuals' expected utility of this contract can be found using (8). Maximizing $\overline{u}(z, k)$ such that the equation above is satisfied gives the perfect competition contract at that price of care; demand for care is found using the definition of demand. The entire perfect competition demand curve can be found by varying $\underline{p}$ and repeating this analysis.

The equilibrium outcome is found as the intersection of the demand curve and the pseudo-supply curve and thus depends on the assumptions on bargaining between insurers and providers. For simplicity, I assume that there are at least two providers, so that the threatening equilibrium can be employed, and that insurers either collude in their negotiations or there is only a single benevolent insurer offering the perfect competition contract. Thus the pseudo supply curve is characterized by $p = (1 + \frac{1}{M} - \frac{1}{M \cdot 1})\gamma D = \gamma D$.

F.2 Private Information and Social Planner's Solution

The social planner accounts for the supply-side constraint that $p = \gamma D$ when selecting the utility-maximizing contract to offer to individuals. To find this optimal contract, first solve the definition of demand for care for copay c:

$$c = (y - z) - u^{-1}(u(y - z) - F^{-1}(1 - D)). \tag{58}$$

Next, rearrange the supply constraint and substitute into the zero profit condition:

$$D = \underline{p}/\gamma, \qquad z = (\underline{p} - c)D \Longrightarrow z = (\underline{p} - c)\underline{p}/\gamma = (\underline{p}^2 - \underline{p}c)/\gamma. \tag{59}$$

These can be substituted into the equation for copay above; slight rearranging yields:

$$0 = (y - (\underline{p}^2 - \underline{p}c)/\gamma) - u^{-1}(u(y - (\underline{p}^2 - \underline{p}c)/\gamma) - F^{-1}(1 - \underline{p}/\gamma)) - c. \tag{60}$$

At any price of care $\underline{p}$, this equation has only a single root in c; the corresponding premium can be recovered from the latter half of (59). This is the zero profit contract at that price that yields a level of demand consistent with the providers' constraint. As before, the expected utility of this contract can be found with (8). The social planner's solution is found by maximizing $\overline{u}(z, k)$ such that (60) and (59) are satisfied by varying $\underline{p}$– sliding along the pseudo supply curve. Demand and price can be found using the definition of demand and the first half of (59).

Note that the social planner's solution only provides a single point, not an entire demand curve– the supply side was accounted for when maximizing utility of the demand side! To generate the entire demand curve, we vary γ and find the social planner's contract (and accompanying price and demand) under different slopes of the marginal cost function.

F.3 Observable Needs and Perfect Competition

The perfect competition demand curve is somewhat easier to derive when medical needs are observable. Here, insurance contracts will be complete, with no copay; instead, they specify a threshold t below which individuals cannot purchase care. The relationship between demand and the threshold is simple: $D = 1 - F(t)$. The zero profit condition likewise simplifies to $z = \underline{p}D$ as there is no copay. The choice of t can thus be reframed as a choice over D as if it were the contractible entity. With $F(\cdot)$ specified as exponential and $\Delta u = 0$ with no copay, we can find the expected utility of any zero profit contract as a function of demand by rearranging (8) as:

$$\overline{u}(D) = u(y - \underline{p}D) + \lambda(D(1 - \log(D)) - 1), \qquad \overline{u}'(D) = -(p(y - \underline{p}D)^{-\rho} + \lambda \log(D)). \tag{61}$$

The first derivative of expected contract utility with respect to demand is strictly decreasing, so there is a single root and thus a unique expected utility maximizing contract any given price. The

threshold of the contract can be found by $t = F^{-1}(1 - D)$. Varying $\underline{p}$ and repeating this procedure generates the entire perfect competition demand curve. The intersection with the pseudo-supply curve $\underline{p} = \gamma D$ establishes the perfect competition equilibrium.

F.4 Observable Needs and Social Planner's Solution

As before, the social planner takes account of the provider's constraint that $\underline{p} = \gamma D$ when selecting the utility-maximizing contract. Combined with the zero profit constraint $z = \underline{p}D$, we find that $z = \gamma D^2$. Substituting this expression for the premium into the first term of (61), we find that expected utility as a function of (implicitly chosen) demand is:

$$\overline{u}(D) = u(y - \gamma D^2) + \lambda(D(1 - \log(D)) - 1), \qquad \overline{u}'(D) = -(2\gamma D(y - \gamma D^2)^{-\rho} + \lambda \log(D)). \tag{62}$$

Once again, expected utility is strictly concave in D, yielding a unique utility maximizing zero profit contract that obeys the provider's constraint. The threshold can be recovered as before, and $\underline{p}$ by $\underline{p} = \gamma D$. As with private information, this generates only a single point; the social planner's demand curve can be found by varying γ to find the optimal contract under different marginal costs of producing care.

This page intentionally left blank.

CURRICULUM VITA

Matthew Noel White was born on March 13, 1982, in the bustling metropolis of Nashua, NH, the second largest city in the Granite State. Raised in Nashua and the neighboring town of Hollis until age 18, he matriculated to Cornell University in August 2000. After spending three misguided years careening between majors in the College of Engineering, he eventually finished his undergraduate work in 2005 with the College of Human Ecology, studying Policy Analysis and Management (PAM). He spent two years as a research assistant to Sean Nicholson of PAM before being inspired to pursue an academic career in economics. Rejected by every other university he applied to, Matthew was accepted by and enrolled in the PhD program at Johns Hopkins University, beginning August 2007. After completing the two years of coursework, he pursued applied and theoretical research in health economics under the advisement of Hülya Eraslan and Chris Carroll. During four years of work on his dissertation, he served as a research assistant to Prof. Carroll and a teaching assistant to Louis Maccini (inter alia). On the academic job market in January 2013, he was hired as an Assistant Professor by the department of economics at the University of Delaware (UD), beginning in August 2013. His applied research on health insurance markets continues at UD, augmented by work on computational methods to accelerate the solution of multidimensional dynamic optimization problems. This dissertation, completed in May 2014, fulfills the requirements of his PhD. Matthew's oral defense occurred twenty-five days before he would have been fired for violating his employment contract, validating his high school's nearly unanimous election of him as Class Procrastinator. *Nosce te ipsum.*